How does dialect exposure affect learning to read and spell? An artificial orthography study

Glenn P. Williams
Nikolay Panayotov
Vera Kempe

https://doi.org/10.1037/xge0000778

Williams, G. P., Panayotov, N., & Kempe, V. (2020) 'How does dialect exposure affect learning to read and spell? An artificial orthography study'. *Journal of Experimental Psychology: General*.

How Does Dialect Exposure Affect Learning to Read and Spell?

An Artificial Orthography Study

Glenn P. Williams[1,2]
Nikolay Panayotov[1]
Vera Kempe[1]

[1] Abertay University
[2] University of Sunderland

Correspondence concerning this article should be addressed to Vera Kempe, Division of Psychology, School of Applied Sciences, Abertay University. Dundee. DD1 1HG. Scotland, UK. E-mail: v.kempe@abertay.ac.uk. Pre-registration, data and code are available at https://osf.io/5mtdj/.

Acknowledgements:

The authors gratefully acknowledge funding from The Leverhulme Trust (Grant #RPG_2016- 039) to Vera Kempe. We also would like to thank Richard Morey & E. J. Wagenmakers for helpful statistical advice but remain responsible for any potential flaws in the statistical analyses.

Abstract

Correlational studies have demonstrated detrimental effects of exposure to a mismatch between a non-standard dialect at home and a mainstream variety at school on children's literacy skills. However, dialect exposure often is confounded with reduced home literacy, negative teacher expectation and more limited educational opportunities. To provide proof of concept for a possible causal relationship between variety mismatch and literacy skills, we taught adult learners to read and spell an artificial language with or without dialect variants using an artificial orthography. In three experiments, we confirmed earlier findings that reading is more error-prone for contrastive words, i.e. words for which different variants exist in the input, especially when learners also acquire the joint meanings of these competing variants. Despite this contrastive deficit, no detriment from variety mismatch emerged for reading and spelling of untrained words, a task equivalent to non-word reading tests routinely administered to young school children. With longer training, we even found a benefit from variety mismatch on reading and spelling of untrained words. We suggest that such a dialect benefit in literacy learning can arise when competition between different variants leads learners to favour phonologically mediated decoding. Our findings should help to assuage educators' concerns about detrimental effects of linguistic diversity.

Keywords: literacy, dialect, artificial language learning

Word count: 14,425

Introduction

In 2013, the BBC reported that a Head Teacher in England had banned use of the local dialect in his Primary School (BBC News, 2013). This decision appears to have been motivated by the notion that dialect exposure creates confusion when beginning readers encounter different variants associated with the same meaning and have to resolve the competition between them. However, direct empirical support for the notion that such competition slows the acquisition of literacy skills is lacking. The aim of the present study is to put this notion to a rigorously controlled test.

Although linguistic diversity is a ubiquitous feature of many languages, most research on how exposure to different varieties affects literacy acquisition has been conducted on minority dialects of English spoken in the United States. A considerable body of evidence has implicated exposure to these minority dialects, and especially the degree of 'dialect density', i.e. the frequency of oral dialect use, as risk factors for reading difficulties (e.g. Charity, Scarborough & Griffin, 2004; Terry, Connor, Johnson et al., 2016, Washington, Branum-Martin, Sun & Lee-James, 2018). Although such a link has not been observed consistently, owing to methodological flaws in the measurement of children's dialect exposure and literacy outcomes in earlier studies (Harber, 1977; Steffensen et al., 1982), the persistent literacy achievement gap in US minority children sustained interest in studying exposure to non-standard varieties. A recent meta-analysis by Gatlin & Wanzek (2015) concluded that there was a moderate negative relationship between exposure to, and use of, non-mainstream American English and literacy outcomes in the absence of significant effects of socio-economic status (SES). While SES has implications for a host of variables such as quality of input, home literacy, attitudes towards literacy, educational provision and teacher expectation, the independent effect of these variables is difficult to control in correlational studies (Artiles et al., 2010). For example, one prominent US minority dialect, African-American English (AAE), has diverged from Mainstream American English (MAE) as a function of, among other things, social and cultural segregation leading to divergent attitudes towards literacy (Labov, 1995). It is therefore important to understand whether dialect exposure exerts a detrimental effect via those socio-cultural and environmental variables or whether it plays a direct causal role in the impairment of emerging literacy.

According to the Linguistic Mismatch Hypothesis (Labov, 1995), dialect exposure increases the mismatch between orthographic and phonological forms thus rendering the discovery of phonologically mediated decoding principles more challenging. Specifically, dialect variants that deviate strongly from standard words so as to essentially constitute competing lexemes (e.g. Scots "bairn" vs. Standard English "child") have to be acquired in addition to learning to read and spell. When learners attempt to establish links between orthographic and phonological representations as postulated in computational models of reading such as the DRC (Coltheart et al., 2001), the CDP+ (Perry et al., 2007, 2010) or the triangle model (Harm & Seidenberg, 2004; Plaut et al., 1996), the activation of competing phonological representations of words with dialect variants (henceforth: contrastive words)

might lead to interference, which should incur additional processing cost. On the other hand, dialect variants characterised by mainly phonological changes that deviate only slightly from words in the standard language (e.g. Scots "hoose" vs. Standard English "house" or AAE "aks" vs. MAE "ask") add inconsistency to the mapping from print to sound. The resulting increased orthographic inconsistency is likely to make the acquisition of decoding skills via application of phoneme-grapheme conversion rules more difficult, particularly for phonologically less consistent orthographies, such as English, which are difficult to decode even without additional dialect variation (but see J. S. Bowers & Bowers [2017]; [2018]; Rastle [2019] for arguments in favour of benefits from considerable morphological transparency of English spelling).

Alternatively, the Linguistic Awareness/Flexibility Hypothesis suggests that high dialect density is a manifestation of limited meta-linguistic awareness of the social and contextual features that cue the appropriate use of one or the other variety (Terry & Scarborough, 2011). Limited meta-linguistic awareness, especially in the phonological domain, has been linked to poorer decoding and comprehension skills (Ehri et al., 2001). Under this account it is not dialect exposure *per se* that impairs literacy acquisition. Rather, the effect is an indirect one: children who persist with dialect use in settings which presuppose use of the mainstream variety betray a lack of meta-linguistic awareness the manifestations of which in other domains like phonology hinder acquisition of decoding skills. As under the Mismatch hypothesis, this deficit should be especially problematic for the decoding of contrastive words where meta-linguistic awareness of contextual information can indicate which variant should be favoured to resolve the competition. The Mismatch and the Awareness accounts need not be mutually exclusive as the direct effect of dialect exposure might be partially mediated by linguistic awareness (Terry & Scarborough, 2011).

The hypothesis that contrastive words elicit reading difficulties was tested by Brown and colleagues (Brown et al., 2015) with 8-13-year-old children exposed to AAE who were asked to read contrastive and non-contrastive words matched for frequency, length and initial phonemes. The contrastive words typically had dialect variants with reduced consonant clusters. The results showed that the higher these children's usage of AAE, assessed through number of AAE features in a sentence repetition task, the longer their reading latencies for contrastive words. This contrastive deficit was computationally simulated in a neural network which instantiated statistical learning of spelling-sound correspondences. The model was first exposed to repeated mappings of phonological to phonological representations within an attractor network (i.e. a task mimicking learning to speak) before being trained to map orthographic onto phonological representations via a layer of hidden units (i.e. a task mimicking learning to read) while still receiving interleaved blocks of phonological exposure to prevent "catastrophic interference" when switching from one type of input to another. Crucially, when the network was initially exposed to AAE variants for half of the words (i.e. variants comprising dialect-appropriate consonant cluster reductions, consonant drops, substitutions, exchanges and devoicing) and then subsequently was trained to read MAE words (the mismatch condition), the cross-entropy

error remained higher for contrastive compared to non-contrastive words. Brown and colleagues interpreted this finding in analogy to the reading of heterophonic homographs - identical spellings of semantically unrelated words that are pronounced differently like "lead" or "wind", which in the absence of contextual information are more difficult to read compared to non-homographic control words (Gottlob et al., 1999; Jared et al., 2012). As predicted by the Linguistic Awareness/Flexibility hypothesis (Terry & Scarborough, 2011), the contrastive deficit was greatly diminished in a second simulation that instantiated nodes coding explicitly for whether a word belonged to AAE vs. MAE, a feature designed to simulate social cues for use of one or the other variety.

While the Brown et al. (2015) simulation undoubtedly provided important insights into potential mechanisms that might be responsible for the difficulty with reading contrastive words, some crucial components of word representation and literacy learning were absent from the model. As a result, it is not entirely clear whether the contrastive deficit in the neural network arises for the same reasons as it did in beginning readers, even if extralinguistic factors are controlled. Firstly, the network lacked a semantic layer precluding instantiation of semantic representations for individual words. Yet beginning readers tend to know the meanings for most, if not all, of the words presented in early literacy training, and start out by employing phonologically mediated decoding to gradually establish direct associations between the new orthographic code and the existing semantic representations (Castles et al., 2018). There was no mechanism in the Brown et al. model by which different variants could be associated with the same meaning. Instead, in the mismatch condition, the network simply learned more words overall as literacy training added an additional set of MAE words which were phonologically similar to some of the already acquired AAE words. As a result, contrastive words shared many phonemes with other variants in the lexicon while non-contrastive words did not, rendering the contrastive deficit – as mentioned above – akin to inhibition from high-frequency heterophonic homographs. Extrapolating from existing models of interactive activation and competition that try to explain neighbourhood effects we hypothesise that adding a semantic layer should retain or even exacerbate the contrastive deficit as bidirectional links between semantic and lexical representations may reinforce non-linear inhibitory connections on the lexical layer (Chen & Mirman, 2012).

Secondly, neither human participants nor the connectionist model exhibited difficulties with reading non-contrastive words nor an overall reading deficit in the variety mismatch (AAE) condition. If potential detriments due to variety mismatch are mainly driven by processing difficulties with contrastive words then the overall amount of literacy problems associated with dialect exposure would mainly depend on the proportion of contrastive words in the input. Yet the Linguistic Mismatch Hypothesis as formulated by Labov (1995) went beyond confining detrimental effects to contrastive words by suggesting that dialect exposure impairs orthographic decoding skills more generally. Similarly, the Linguistic Awareness/Flexibility Hypothesis (Terry & Scarborough, 2011) also asserts that limited dialect awareness should impair beginning readers' general phonological decoding skills. However, to directly confirm detrimental effects of dialect exposure beyond

contrastive words one would have to test beginning bi-dialectal readers' decoding skills independently of their word knowledge (Castles et al., 2018) and show that their nonword reading skills are impaired compared to learners without dialect exposure. Such a test, which was absent from both the behavioural study and the computational simulation in Brown et al. (2015), will be included in the present study.

Thirdly, beginning readers never learn only to read but also to spell, as primary schools tend to incorporate writing instruction into their curricula from early on (Cutler & Graham, 2008). Spelling training strengthens the connections between individual phonemes and graphemes thereby promoting use of decoding skills. In children, phonological spelling ability has been shown to predict subsequent development not just of spelling but, crucially, reading skills (Caravolas et al., 2001), confirming earlier proposals that in the early stages of literacy learning, phonological spelling ability drives the development of reading (Frith, 1985). For adults learning an artificial script, Taylor et al. (2017) showed that including a spelling task in addition to other tasks emphasising spelling-sound as opposed to spelling-meaning mappings into the training regimen encouraged a phonologically mediated reading acquisition strategy. While the child participants in Brown et al. (2015) would certainly have engaged in spelling practice during their schooling, the neural network did not include bi-directional links that could have instantiated a "spelling path", i.e. a path from phonological to orthographic representations (see Houghton & Zorzi, 2003), and we are not aware of any attempts to computationally model the contribution of spelling practice to emergent reading skills. Yet by promoting explicit reliance on links between graphemes and phonemes as the primary reading strategy (Ellis & Cataldo, 1990) which minimises reliance on direct word retrieval, and, hence, the possibility for lexical competition, spelling training might alleviate potentially detrimental effects of dialect exposure.

The evidence discussed so far was obtained in studies investigating the process of learning to read English, a deep orthography with a fair amount of inconsistent phoneme-grapheme mappings that still is often taught without placing sufficient emphasis on phonological mediation (Castles et al., 2018). For such inconsistent spelling systems, dialect exposure may be particularly detrimental as it can further hinder acquisition of already difficult-to-discover decoding rules. By contrast, learning to decode mappings from sound to spelling is easier in more consistent orthographies, and, consequently, dialect exposure may have less of a detrimental effect. It is even possible that more rapidly acquired decoding skills in consistent orthographies can render the decoding of words of the standard variety unperturbed by the existence of competing dialect variants. To our knowledge, the only more consistent orthography for which the role of dialect exposure in literacy learning has been investigated is German. Bühler and colleagues (Bühler et al., 2018) examined early literacy skills in children exposed to Swiss German dialect and compared them with children exposed only to Standard German either in Switzerland or in Germany. The results showed that dialect exposure was associated with higher preschool literacy-related skills measured by the ability to identify, categorise and synthesise onsets, rimes and individual phonemes, in the absence of differences in SES between the groups. Structural equation modelling revealed that only when preschool literacy-related skills were controlled was

there a negative effect of dialect exposure on Grade 1 literacy skills, which was more pronounced in spelling owing to the fact that German's phoneme-grapheme mappings are less consistent than the grapheme-phoneme mappings. This finding exposes the multiple loci of effects that early dialect exposure might have: On the one hand, benefits from early dialect exposure on literacy-related skills might arise from enhanced sensitivity to phonological variation thereby increasing metalinguistic skills that benefit phonological awareness. On the other hand, residual negative effects of dialect exposure on subsequent literacy acquisition may reflect the consequences of decreased consistency in spelling-sound mappings as well as the difficulty associated with first having to learn a number of new lexemes in order to master literacy in the standard language. This suggests that in orthographies with greater feed-forward (reading) consistency, potentially detrimental effects of dialect exposure may be offset by its contribution to enhanced phonological awareness which, in turn, can aid phonological mediation of literacy learning.

To gain further clarity, our study asked whether there is a causal relationship between variety mismatch and difficulties with acquiring decoding skills when confounding extra-linguistic variables that may impact the acquisition of these skills are controlled. By variety mismatch we mean a situation where another variety (e.g. a regional dialect) is used outside of the context of literacy acquisition. To achieve this control, we employed an artificial language learning paradigm combined with an invented script, a methodology that has successfully been used to explore various factors that affect the early stages of learning to read (e.g. Taylor et al., 2017; Taylor, Plunkett, & Nation, 2011; for a review see Vidal et al., 2017). We attempted a conceptual replication of the contrastive deficit demonstrated in Brown et al. (2015) to confirm whether variety mismatch is indeed the cause of deficits associated with dialect exposure. Crucially, we also asked whether variety mismatch affects general decoding skills as assessed via reading of untrained words. Here, we perform these tests with adult learners to provide a baseline for future comparison with children. We seek to provide proof of concept for how dialect exposure *per se* can affect literacy learning under optimal learning conditions associated with a mature cognitive system: Detrimental effects in adults would suggest that dialect exposure is bound to hinder literacy learning by virtue of increasing the amount of interference in the input, and detrimental effects in children may be inevitable. However, if no detrimental effects are observed in adults then detrimental effects in children may arise from how dialect exposure interacts with a less mature cognitive system or due to confounding factors that affect children who are exposed to dialects.

The Present Study

We report three experiments designed to investigate effects of dialect exposure on the acquisition of decoding skills in inconsistent and consistent orthographies. For the present study, we defined dialect exposure following Brown et al. (2015) as exposure to variants that entail phonological, but not lexical, changes (e.g. English "house" vs. Scots "hoose" or MAE "ask" vs. AAE "aks"). The focus on phonological variants was motivated by ecological validity based on a corpus analysis (the Gruffalo-corpus described below) of

a range of Scottish English dialects (Johnston, 2007). Effects of exposure to lexical variants (e.g. English "children" vs. Scots "bairns") are beyond the scope of the current study. Below we briefly preview the rationale behind the three experiments.

We start by reporting a conceptual replication of Simulation 1 in Brown et al. (2015) that examines effects of dialect exposure on learning to read an inconsistent artificial orthography[1](Experiment 1). To explore the role of semantic information we compared performance for words presented with and without accompanying pictures. Availability of semantic information was crossed with dialect exposure: In the Variety Match conditions participants encountered the same words during initial exposure and during reading training. In the Variety Mismatch conditions half of the words underwent phonological changes between exposure and reading training to loosely resemble a situation in which learners initially are exposed to a dialect at home before being introduced to the standard variety at school. However, because Experiment 1 was conceived as a replication of Brown et al. (2015), it did not include spelling training and did not vary orthographic consistency, two factors which need to be considered to be able to generalise cross-linguistically. To address these two limitations, we compared Variety Match and Mismatch conditions on learning to read and spell a consistent (Experiment 2a) and an inconsistent (Experiment 2b) orthography. We found that learning to read and spell an entirely unfamiliar inconsistent artificial orthography proved to be a very difficult task that may require more extensive training. In Experiment 3 we therefore replicated Experiment 2b with a longer training phase, a larger sample size and semantic information throughout. All experiments received ethical approval from Abertay University's Ethics Committee and were programmed to be compatible with all desktops and Android systems using most web browsers.

Experiment 1: Effect of variety mismatch on learning to read an inconsistent orthography.

Method

Participants. One hundred and twelve participants (aged 14 – 67, M = 31.46, SD = 11.09, with 68 self-reported as female, 43 self-reported as male, and 1 self-reported as other)[2] were recruited from the crowdsourcing website Prolific Academic. All participants reported English as their native language, and no known mild cognitive impairments or dementia, and were reimbursed £4.20. Participants' mean proficiency in English on a 1-5 Likert scale was 4.85 (SD = 0.59, range 1 [elementary] – 5 [native or native-like]). Despite declaring English as native language, ten participants rated their English proficiency as below 5. Sixty-two participants reported knowing only English while 50 participants also knew French (listed 29 times), Spanish (listed 22 times), German (listed 10 times) and 29

[1] Experiment 1 was conducted last but is reported first to maintain the logic of reporting first the replication attempt of the connectionist simulation reported in Brown et al. (2015).

[2] One participant self-reported an age of 14 which we assume is a typo as Prolific Academic enforces a minimum age of 18.

other languages (listed a total of 44 times). Only eight participants were familiar with logographic scripts. An additional 7 participants were tested but not included either because they gave the same response on all trials, their responses on most trials repeated the previous trial, their responses were in English rather than in the artificial language or were inaudible, or because a technical difficulty had occurred (e.g. losing trials due to poor internet connection), and when recruitment inadvertently extended beyond our pre-registered cut-offs for a given list.

Materials.

Grapheme and Phoneme Inventory. We generated thirteen graphemes (for a list of graphemes and criteria for inclusion see Appendix A) consisting of two to four curved or straight strokes as common to most alphabetic writing systems (Changizi & Shimojo, 2005). The phoneme inventory consisted of eight consonants [m], [n], [s], [k], [b], [d], [f] and [l] as well as the five cardinal vowels [ɑ], [ɛ], [i], [ɔ], [u]. Additionally, the dialect phonemic inventory included an additional phoneme, [x], which replaced [k] in certain contexts, as described below.

Words. Using this phoneme inventory, we constructed 42 artificial words distributed across six syllabic templates (3 monosyllabic, 3 bisyllabic) adhering to constraints of English phonotactics (Crystal, 2003; Harley, 2006). To constrain phonological complexity and to avoid overly predictive clusters, words contained no more than one consonant cluster and no cluster with more than two consonants. Applying these rules to a string generation algorithm (accessible at https://osf.io/5mtdj/), we produced all possible phoneme permutations per syllable template, and selected seven strings from each template type, by removing strings with phoneme repetitions and ensuring a similar distribution of phonemes across items. To capture a range of English neighborhood densities, we selected a roughly equal number of words with high and low phonological neighbourhood densities according to the Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities (Marian et al., 2012) database using the total neighbour metric (i.e. including substitutions, additions, and deletions) resulting in a mean neighbourhood density of 2.88. To minimise confusability of words, our final list was filtered such that each word differed from each other word by a length-normalised Levenshtein Edit Distance (nLED)[3] of at least 0.5, resulting in an average nLED of 0.86, ensuring sufficient variability across items. This restriction was applied as variability has been shown to support learning of grapheme-phoneme-correspondences (Apfelbaum et al., 2013). Thirty words were presented during exposure and literacy training, while twelve words (two words from each syllable template) were retained for testing only (henceforth: untrained words). All words are listed in Appendix C.

[3] A widely used normalised measure computed by dividing the number of insertions, deletions, and substitutions required to transform one string into another by the larger of the two string lengths (Levenshtein, 1966).

Words and isolated phonemes were recorded by a male and a female speaker in a soundproof booth with a Zoom H4n audio recorder, using normal prosody with stress on the first syllable in the bi-syllabic items. Speaker voice was counterbalanced across participants. Sound files were normalised, with noise filtered using the Audacity audio suite (Mazzoni & Dannenberg, 2016) and extraneous silences were trimmed using Praat (Boersma & Weenik, 2017). In the Picture condition, words were randomly combined with images taken from the revised Snodgrass and Vanderwart image set of colourised images provided by Rossion and Pourtois (2004). For specifics of picture selection see Appendix D.

Orthography. To create orthographic inconsistencies, we introduced two conditional rules to supplement one-to-one mappings of graphemes to phonemes. First, the phoneme /l/ was rendered by its corresponding grapheme in all contexts except in five instances when it was preceded by /b/ or /s/, in which case it was spelled using the grapheme otherwise assigned to /n/ so that, for example, /blaf/ was spelled as the artificial equivalent of *BNAF*. Second, the phoneme /s/ was rendered by its corresponding grapheme in all contexts except in five instances when it was preceded by /n/ in which case it was rendered by the grapheme otherwise assigned to /f/ so that, for example, /snid/ was spelled *FNID*. It is important to emphasise that these conditional rules introduced a roughly similar amount of inconsistency in both directions: In terms of feed-forward consistency (spelling-sound correspondences required for reading), the artificial grapheme signifying *F* was pronounced as /s/ 27% of times and as /f/ 73% of times. Similarly, the artificial grapheme signifying *N* was pronounced as /n/ 67% of times and as /l/ 33% of times. In terms of feed-back consistency (sound-spelling correspondences required for writing), the phoneme /s/ was spelled as (the artificial equivalent of) the letter *S* 74% of times and as *F* 26% of times. Similarly, the phoneme /l/ was spelled as *L* 75% of times and as *N* 25% of times. These conditional spelling rules were matched across word type resulting in five contrastive and five non-contrastive words with irregular spelling. For one contrastive and one non-contrastive word, both conditional spelling rules applied simultaneously so that /sloku/ and /slinab/ were spelled as *FNOKU* and *FNINAB*, respectively.

Simulating Dialect Exposure based on the Gruffalo-corpus. Because processing of phonological vs. lexical variation might rely on different mechanisms as discussed above, we restricted this study to just one type of variation, namely that which is most frequent in prominent naturally occurring dialect varieties of English like Scots. This determination requires frequency estimates from transcribed corpora of dialect use which, to our knowledge, do not exist. We therefore consulted translations of the two popular children's books "The Gruffalo" and "The Gruffalo's Child" (Donaldson, 1999, 2005), written in Standard British English (SBE), into a number of varieties of Scots, including Dundonian, Glaswegian, and Doric, to obtain such a dialect corpus. The seven books comprising this corpus are listed in Appendix B. This approach, in essence, amounts to treating the translators of the original version of "The Gruffalo" as native dialect informants. Using a corpus derived from children's verses gives us estimates for how dialects differ from

standard varieties for linguistic content that is appropriate for the age group at which literacy is acquired.

The Gruffalo corpus comprised 310 translated word types. Each of the Scots words in each Gruffalo translation was aligned with its SBE equivalent and coded for whether it differed lexically resulting in a Scots word not existent in SBE (e.g. *big – muckle*[4]) or phonologically (e.g. *mouse – moose*). To validate this categorisation we computed nLEDs between the SBE and Scots variants for each category (lexical variants: $M = 0.80$, phonological variants: $M = 0.40$). Phonological differences were further sub-categorised as phoneme drops (e.g. *and – an*), substitutions (e.g. *bright – bricht*), or insertions (e.g. *it's – hit's*), and whether diphthongisation (e.g. *ahead – ahaid*) or monophthongisation (e.g. *mouse – moose*) occurred[5]. A total of twenty-six words involved a difference which could not be reliably categorised as lexical or phonological. Words that arose from paraphrasing the SBE phrases (e.g. *"...that no Gruffalo should ever set foot" – "it wid come tae nae guid if..."*) were excluded from our analysis.

Analysing the distribution of variants revealed that 93.23% of word types and 53.01% of word tokens were contrastive, i.e. had a dialect variant. Of these contrastive words, 49.48% of types and 63.94% of tokens had variants with phonological differences, confirming that phonological variation was indeed the most common variation. Of the phonological variants, the most frequent ones were phoneme substitutions (79.91% of all phonological variant tokens) and consonant drops (24.87% of all phonological variant tokens)[6]. These estimates suggest that inclusion of 50% of words with dialect variants as in Brown et al. (2015) provides an ecologically valid amount of dialect variation. We therefore implemented a range of variations that mimicked those found in the Gruffalo-corpus as listed below:

(a) *consonant substitution*: [k] was changed to [x] in all positions (e.g. /skub/ changed to /sxub/).

(b) *consonant drop*: [d] was dropped in final position (e.g. /snid/ changed to /sni/).

(c) *vowel change*: [ɛ] and [ɑ] were replaced with [i] and [ɔ], respectively (e.g. /nɛf/ changed to /nif/) and /nɑl/ changed to /nɔl/) in all positions. In instances where multiple changes

[4] Described by the Dictionary of the Scots Language (n.d.) as an adjective meaning 'of size or bulk: large, big, great'.

[5] In some cases, such as the Scots *ken* [*know]*, this change is recorded as a lexical change as the phonology of the word changes dramatically, i.e. [kɛn] from [nəʊ] despite maintaining the root of the word. Moreover, due to lack of standardisation of Scots spelling, we only could include phonological changes that were orthographically rendered. For example, while the voiceless velar fricative was orthographically rendered in some cases, /x/ (e.g. *right – richt*), in others it was not (e.g. *loch – loch*), and these changes could not be counted in the corpus analysis.

[6] Note that these values sum to more than 100% as words could include both phoneme substitutions and consonant drops, amongst other changes.

could apply all were implemented in the "dialect" so that, for example, /skɛfi/ became /sxifi/ and /flɛsɔd/ becomes /flisɔ/.

Procedure.

Participants were instructed that they would learn to read a "made-up" language and that it was important to perform the task in a quiet environment and to not take any notes during participation. To ensure compliance the instructions misled participants into believing that detection of cheating on our part could jeopardise reward. After receiving instructions describing the experimental procedure and providing consent compatible with the General Data Protection Regulation, participants were asked to check the working order of their microphone and headphones/speakers. The experiment consisted of three components: exposure, training and testing (see Table 1). In the first exposure block, participants heard all thirty training words one by one in randomised order. They then viewed each grapheme one by one (cycling twice during the set), accompanied by the sound of the isolated phoneme. Crucially, phonemes were randomly assigned to graphemes for each participant to reduce the potential impact of any systematic differences in accessibility of grapheme-phoneme pairs. Following recommendations to include time limits preventing participants from taking notes in learning experiments (Rodd, 2019), each grapheme disappeared after 1,000 ms. This process was repeated once, exposing participants to each grapheme-phoneme combination for a total of four times.

Next, participants proceeded to the reading training, which was interleaved with more exposure. To this end, the set of thirty words was randomly split into three reading training blocks of ten words each. For each item, participants saw a string of graphemes and had to read the target word out loud. To avoid recording long silences we timed participants' responses by presenting a moving hand in a clock indicating the onset, duration and offset of the 2500ms recording window. In the Picture condition, orthographic representations were always accompanied by pictures to simulate availability of semantic context. Although script is typically not accompanied by pictures, we deemed such a procedure justified given that reading rarely is context-free and confined to single words and children's early reading materials frequently contain illustrations. Upon completion of each recording, participants received auditory feedback by listening to the target sound form. Each ten-item training block was presented twice in a row to equate number of exposures per word with Experiment 2 to maintain comparability (for an overview of the task sequences in all three experiments see Table 1). The first such block was followed by another exposure to all thirty words before proceeding to the second block of training, followed by another exposure. In total, participants were exposed to the set of 30 words three times – once at the beginning, once after the first, and once after the second reading training block. After completing the third reading training block, participants were tested on reading of the thirty trained and the twelve untrained words, all presented in random order without auditory feedback.

Table 1: Task sequence in all experiments. Randomly presented word numbers are given in parentheses to indicate number of words per task. [1,2] - Counterbalanced order of reading and spelling training [3] - Prior to block 4, words were re-randomised and re-partitioned.

	Experiment 1	**Experiment 2**		**Experiment 3**
	inconsistent	**2a - consistent**	**2b - inconsistent**	**inconsistent**
training block 1	exposure (1-30) grapheme learning reading (1-10) reading (1-10)	exposure (1-30) grapheme learning reading[1,2] (1-10) spelling[2,1] (1-10)	exposure (1-30) grapheme learning reading[1,2] (1-10) spelling[2,1] (1-10)	exposure (1-30) grapheme learning reading[1,2] (1-10) spelling[2,1] (1-10)
training block 2	exposure (1-30) reading (11-20) reading (11-20)	exposure (1-30) reading[1,2] (11-20) spelling[2,1] (11-20)	exposure (1-30) reading[1,2] (11-20) spelling[2,1] (11-20)	exposure (1-30) reading[1,2] (11-20) spelling[2,1] (11-20)
training block 3	exposure (1-30) reading (21-30) reading (21-30)	exposure (1-30) reading[1,2] (21-30) spelling[2,1] (21-30)	exposure (1-30) reading[1,2] (21-30) spelling[2,1] (21-30)	exposure (1-30) reading[1,2] (21-30) spelling[2,1] (21-30)
training block 4	n/a	n/a	n/a	exposure[3] (1-30) reading[1,2] (1-10) spelling[2,1] (1-10)
training block 5	n/a	n/a	n/a	exposure (1-30) reading[1,2] (11-20) spelling[2,1] (11-20)
training block 6	n/a	n/a	n/a	exposure (1-30) reading[1,2] (21-30) spelling[2,1] (21-30)
testing	reading (1-42)	reading[1,2] (1-42) spelling[2,1] (1-42)	reading[1,2] (1-42) spelling[2,1] (1-42)	reading[1,2] (1-42) spelling[2,1] (1-42)

Crucially, in the Variety Mismatch condition, participants heard the dialect variants of contrastive words during all exposure blocks but were presented with the standard variants during reading feedback. The source code for the experiment can be found at https://osf.io/5mtdj/. The mean completion time was 52.43 minutes ($SD = 25.27$). To ensure equal number of participants per condition a randomised sequence of eight conditions comprising a crossing of Variety condition, Picture condition and Speaker Voice (female vs. male) was created and administered consecutively over the course of about two weeks thereby ensuring pseudo-random assignment to all conditions. Repeated sign-up of participants was blocked by the crowdsourcing website.

Data analysis. We used R (Version 3.5.2; R Core Team, 2018) and the R-packages *brms* (Version 2.7.0; Bürkner, 2017, 2018), *broom.mixed* (Version 0.2.2; Bolker & Robinson, n.d.), *emmeans* (Version 1.3.2; Lenth, 2019), *english* (Version 1.2.0; Fox et al., 2019), *here* (Version 0.1; Müller, 2017), *irr* (Version 0.84.1; Gamer et al., 2019), *kableExtra* (Version 1.0.1; Zhu, 2019), *knitr* (Version 1.22; Xie, 2015), *lme4* (Version 1.1.20; Bates et al., 2015), *lmerTest* (Version 3.1.0; Kuznetsova et al., 2017), *papaja*

(Version 0.1.0.9842; Aust & Barth, 2018), and *tidyverse* (Version 1.2.1; Wickham, 2017) for data preparation, analysis, and presentation. All data processing and analyses were prereqistered and are hosted on the Open Science Framework (https://osf.io/5mtdj/). Any deviations from our pre-registered analysis plan are outlined and justified in the pre-registration deviations documents.

Results

Coding. Two coders (GPW and VK) transcribed all reading responses while blind to each participant's condition. A coding convention was adopted for the 13 target phonemes in the artificial language using the CPSAMPA (Marian et al., 2012) simplified notation of IPA characters such that [ɑɛiɔu] became a, E, i, O, u while the consonants were coded using the letters m, n, s, k, b, d, f, l and x. All extraneous, i.e. non-target phonemes were rendered by single Latin characters that provided the closest match so as to be able to compute nLEDs, which constitute a more gradual and fine-grained performance measure than error rates, allowing us to distinguish near-matches from entirely erroneous productions akin to cross-entropy errors in the neural network simulation by Brown et al. (2015). We computed inter-coder reliability by obtaining intra-class correlations between the two coders' nLEDs, using the *irr* R-package (Gamer et al., 2019). We used a single-score, absolute agreement, two-way random effects model based on the summed nLEDs for each participant. Inter-coder reliability was $F(111.00,111.86) = 27.71$, $p < .001$, ICC = 0.931 [95% CI = 0.901; 0.952]. The 95% confidence interval around the parameter estimate indicates that the ICC falls above the bound of .90, which suggests excellent reliability across coders (Koo & Li, 2016). However, in instances of discrepancy between the coders we based further analyses on the smaller of the two nLEDs thereby adopting a lenient coding criterion based on the assumption that a participant response counts as acceptable if at least one of the coders can match it to the target as closely as possible.

Model Fitting. We performed separate analyses for the training and testing phases. Our dependent variable, the leniently coded nLED, was arcsine square root transformed to adjust for the bounded nature of values between 0 and 1. We performed frequentist and Bayesian analyses. Bayesian analyses, although not fully adopted as standard in studies of this kind, provide a range of additional advantages (Nicenboim & Vasishth, 2016; Vasishth et al., 2018): Maximal random effect structures (Barr et al., 2013) can be fitted without convergence problems and data can be interrogated directly for null-effects. In our description and interpretation of the results we will focus on those effects that reached significance in the frequentist analysis and had credible intervals that did not include 0 in the Bayesian analysis (marked in boldface in all tables).

For the frequentist analyses, we modelled the data with linear mixed effects models fitted using the *lme4* R-package (Bates et al., 2015). Statistical significance of each term was evaluated with *p*-values approximated using the Satterthwaite method implemented in the *lmerTest* R-package (Kuznetsova et al., 2017). We used the maximal random effects structure that allowed for model convergence throughout (Barr et al., 2013).

For the Bayesian analyses, we fitted linear mixed-effects models using the *brms* R-package (Bürkner, 2017, 2018) with the same fixed effects as in the frequentist models and a maximal random effects structure. To simplify the definition of priors for the estimated parameters, we scaled and centred the dependent variable on a mean of 0 with a standard deviation of 1. We used a regularising, weakly informative prior, *Normal*(0*,* 1), for the intercept term. Additionally, we used an informative prior for all fixed effects terms, defined as *Normal*(0*,* 0.2), except for fixed effects involving time-terms. This prior places a larger probability on small effects for the parameter estimates. For fixed effects including time terms (i.e. each time term and any interactions of other effects with time terms in the training phase only), we used very weakly informative priors, defined as *Normal*(0*,* 10), which allows these effects to be dominated by the likelihood. We also used regularising priors for the correlation parameters, *LKJ*(2), which down-weights perfect correlations (Vasishth et al., 2018). Additionally, the standard deviations of random effects and the residual error used the default priors in *brms* at the time of writing, which are defined as half Student's-*t* priors (i.e. constrained to be non-negative) with 3 degrees of freedom and, minimally, a scale parameter of 10. Without a predefined region of practical equivalence (Kruschke & Liddell, 2018), we used the 95% credible interval around the posterior mean to summarise these models. As Nicenboim and Vasishth (2016) note, the 95% credible interval provides the range of values within which the true value of the parameter lies with 95% probability given the model and data. Thus, when a 95% credible interval includes zero, we conclude that we do not have sufficient evidence against a null result. However, when a 95% credible interval does not include zero, we conclude that we have evidence for a non-zero directional effect (see Bürkner & Vuorre, 2019 for use of similar criteria)[7].

Training. Training data were modelled using growth-curve analyses (e.g. Mirman, 2014) to establish change in performance over time, i.e. from block to block, across conditions. Time was modelled using fixed effects of orthogonal linear and quadratic polynomials to capture the potential non-linear change in performance over the six half-blocks of ten words as learning progressed. Because interactions of quadratic terms with other fixed effects do not lend themselves to meaningful interpretation, they will not be considered further. The model also included sum-coded fixed effects of Picture condition (picture vs. no picture), Variety (match vs. mismatch) and Word Type (contrastive vs. non-contrastive). We used nested fixed effects for these terms (see Schad et al., 2018 for a discussion of this approach), with Word Type nested within the interaction between all other fixed effects. As a result of this parameterisation, the intercept represents the average of condition means throughout the entire time window (and not at the first block), individual terms (except Word Type) represent main effects for the given term, and Word Type effects represent simple effects within each combination of the other factors (e.g. Word Type

[7] We initially attempted to evaluate evidence for and against the null hypothesis for each term in our model using Bayes factors calculated using the generalTestBF function from the *BayesFactor* R-package (Morey & Rouder, 2018). However, this resulted in Bayes factors with a large proportional error. Following this, we calculated Bayes factors using the hypothesis function from the *brms* R-package (using the Savage-Dickey density ratio). However, as Nicenboim and Vasishth (2016) discuss, with wide, weakly informative priors, the Bayes factor will always favour the null hypothesis as the alternative hypothesis is penalised for including large (and unlikely) values in the prior. We there rely on the 95% credible interval, rather than Bayes factors, to interpret non-significant results.

within each level of Picture and Variety conditions over the entire time). All other interactions aside from those involving Word Type are interpreted as usual. In the frequentist model, the random effect structure included zero-correlation random intercepts and slopes of Picture condition, Variety condition, and their interaction by items, and random intercepts, slopes (including correlations) for the linear and quadratic Time terms, Word Type, and their interaction by participants. The results of the models including parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 2.

Table 2: Parameter estimates for the models fitted to nLEDs from the training phase in Experiment 1. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Block (B) = 1 – 6, Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture condition (PC) = picture vs. no picture (P vs. NP), Word Type (WT) = contrastive vs. non-contrastive.

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.72	0.03	[0.65, 0.78]	21.52	< .001	0.01	0.07	[-0.13, 0.15]
Block	**-7.81**	**0.63**	**[-9.05, -6.57]**	**-12.38**	**< .001**	**-0.48**	**0.04**	**[-0.56, -0.40]**
Block^2	**1.67**	**0.49**	**[0.71, 2.64]**	**3.40**	**.001**	**0.10**	**0.03**	**[0.04, 0.16]**
Picture Condition	0.02	0.03	[-0.04, 0.08]	0.74	.463	0.04	0.06	[-0.08, 0.17]
Variety Condition	-0.02	0.03	[-0.08, 0.04]	-0.69	.494	-0.04	0.06	[-0.16, 0.08]
B × PC	0.29	0.63	[-0.95, 1.52]	0.46	.650	0.01	0.04	[-0.07, 0.09]
B^2 × PC	-0.46	0.49	[-1.42, 0.50]	-0.94	.351	-0.03	0.03	[-0.09, 0.03]
B × VC	-1.19	0.63	[-2.43, 0.05]	-1.89	.062	-0.07	0.04	[-0.15, 0.01]
B^2 × VC	-0.32	0.49	[-1.29, 0.64]	-0.66	.512	-0.02	0.03	[-0.08, 0.04]
PC × VC	0.00	0.03	[-0.06, 0.06]	0.13	.899	0.00	0.06	[-0.13, 0.12]
B × PC × VC	**1.29**	**0.63**	**[0.05, 2.52]**	**2.04**	**.044**	**0.08**	**0.04**	**[0.00, 0.16]**

B^2 × PC × VC	-0.40	0.49	[-1.36, 0.56]	-0.81	.418	-0.02	0.03	[-0.08, 0.03]
NP, VMis: WT	**-0.04**	**0.02**	**[-0.07, -0.00]**	**-2.21**	**.031**	**-0.07**	**0.04**	**[-0.14, -0.00]**
P, VMis: WT	-0.03	0.02	[-0.06, 0.00]	-1.80	.077	-0.06	0.04	[-0.13, 0.01]
NP, VMa: WT	-0.01	0.02	[-0.04, 0.03]	-0.45	.655	-0.01	0.03	[-0.08, 0.05]
P, VMa: WT	-0.01	0.02	[-0.05, 0.02]	-0.80	.429	-0.03	0.04	[-0.09, 0.05]
B, NP, VMis: WT	-0.23	0.77	[-1.74, 1.29]	-0.29	.769	-0.01	0.05	[-0.11, 0.08]
B^2, NP, VMis: WT	0.12	0.76	[-1.36, 1.61]	0.16	.871	-0.11	0.05	[-0.20, -0.02]
B, P, VMis: WT	-1.72	0.77	[-3.23, -0.21]	-2.24	.027	-0.03	0.05	[-0.12, 0.07]
B^2, P, VMis: WT	0.89	0.76	[-0.59, 2.37]	1.17	.242	-0.01	0.05	[-0.10, 0.09]
B, NP, VMa: WT	-0.44	0.77	[-1.95, 1.08]	-0.56	.573	0.00	0.05	[-0.08, 0.10]
B^2, NP, VMa: WT	0.73	0.76	[-0.75, 2.21]	0.96	.337	0.06	0.05	[-0.04, 0.15]
B, P, VMa: WT	-0.10	0.78	[-1.62, 1.43]	-0.12	.902	0.05	0.04	[-0.04, 0.13]
B^2, P, VMa: WT	**1.62**	**0.76**	**[0.12, 3.11]**	**2.12**	**.035**	**0.10**	**0.05**	**[0.01, 0.19]**

Our results show reduction of nLEDs across over six half-blocks of training indicating that participants' reading of the artificial script improved over time. The significant quadratic term suggests that in many instances more progress was made between blocks 1 and 2 than between blocks 2 and 3. Crucially, the contrastive deficit was significant in the Variety Mismatch condition without pictures and marginally significant in the Variety Mismatch condition with pictures, broadly confirming greater difficulties with reading contrastive words (see Figure 1). In addition, there was a significant 3-way interaction between Block, Picture condition and Variety condition, which, however, is not of interest to the main questions of this study.

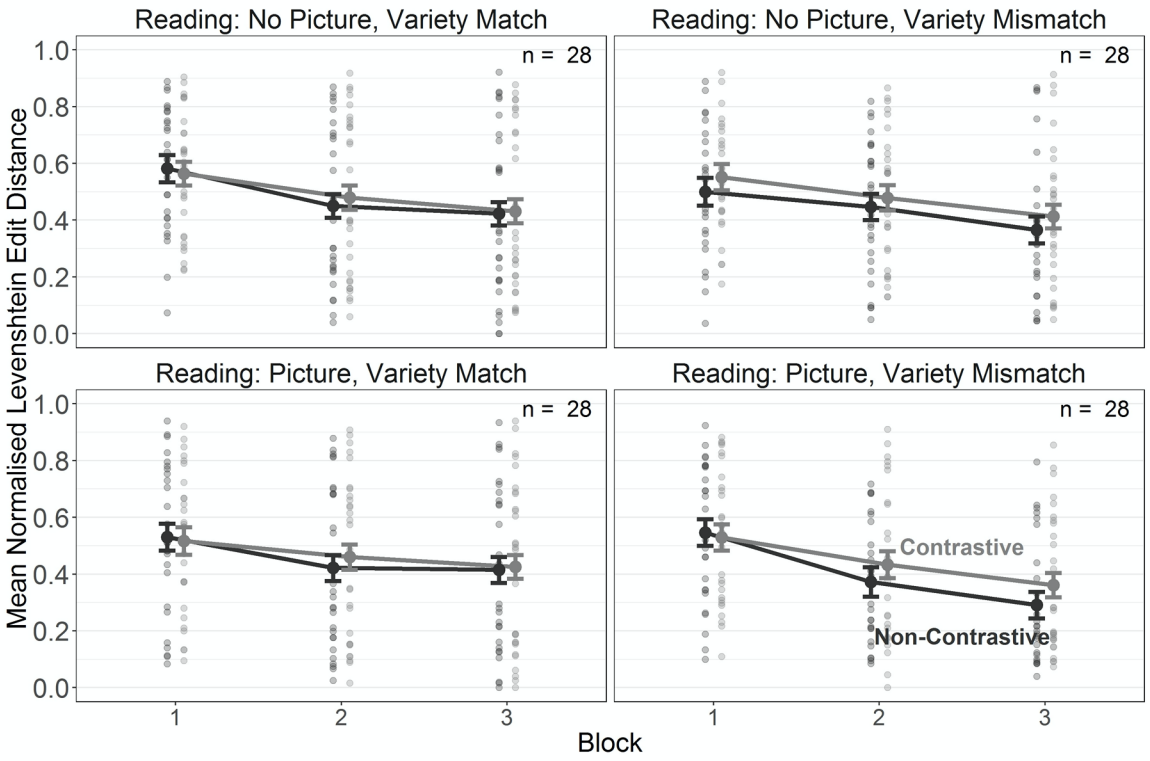

Figure 1: nLEDs for reading training of contrastive and non-contrastive words over 3 training blocks (coded as 6 half-blocks in the analyses but presented as 3 blocks for comparability with Experiment 2) in the Variety Match and Mismatch conditions in Experiment 1. Error bars indicate ± 1 *SE* of the mean.

Testing. For the analysis of the testing phase, we used the same fixed effect structure as for the analysis of the training phase with the exclusion of the linear and quadratic effects of Block. The only difference was that here Word Type was modelled using Helmert contrasts, such that contrastive words were compared to non-contrastive words and untrained words were compared to the average of contrastive and non-contrastive words (i.e. the trained words). For the testing phase, the random effects structure included random intercepts and slopes of Picture condition, Variety condition, and their interaction by items, and random intercepts and slopes of Word Type by participants. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 3.

Table 3: Parameter estimates for the models fitted to nLEDs from the testing phase. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0) in Experiment 1. Values of 0 with a sign indicate the direction of the estimate before rounding. Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture condition (PC) = picture vs. no picture (P vs. NP), Word Type (WT) = contrastive vs. non-contrastive, Word Familiarity (WF) = familiar vs. unfamiliar (novel)

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.61	0.04	[0.54, 0.68]	16.15	< .001	0.01	0.08	[-0.14, 0.19]
Picture Condition	0.04	0.04	[-0.03, 0.11]	1.06	.292	0.06	0.07	[-0.08, 0.20]
Variety Condition	-0.05	0.04	[-0.12, 0.02]	-1.33	.186	-0.08	0.07	[-0.21, 0.07]
PC × VC	0.02	0.04	[-0.05, 0.09]	0.64	.524	0.04	0.07	[-0.09, 0.18]
NP, VMis: WT	-0.02	0.02	[-0.05, 0.01]	-1.13	.262	-0.03	0.03	[-0.10, 0.03]
P, VMis: WT	-0.03	0.02	[-0.07, -0.00]	-2.00	.052	-0.07	0.03	[-0.13, -0.00]
NP, VMa: WT	0.00	0.02	[-0.04, 0.04]	0.02	.985	0.00	0.03	[-0.07, 0.07]
P, VMa: WT	0.00	0.02	[-0.03, 0.03]	-0.23	.817	-0.01	0.03	[-0.07, 0.06]
NP, VMis, WF	0.01	0.01	[-0.02, 0.04]	0.74	.458	0.02	0.03	[-0.04, 0.08]
P, VMis, WF	0.03	0.01	[-0.00, 0.06]	1.94	.055	0.06	0.03	[0.00, 0.12]
NP, VMa, WF	0.01	0.02	[-0.03, 0.04]	0.32	.749	0.01	0.03	[-0.05, 0.07]
P, VMa, WF	**0.04**	**0.01**	**[0.01, 0.07]**	**2.85**	**.005**	**0.08**	**0.03**	**[0.02, 0.14]**

We found that the contrastive deficit failed to reach significance in the Variety Mismatch condition. The effect of Word Familiarity was significant in the Variety Match condition with pictures and fell short of significance in the Variety Mismatch condition with pictures suggesting that participants were able to capitalise on knowledge of the phonological form of trained items either by using partial phonological decoding or direct access from the depicted meaning (see Figure 2).

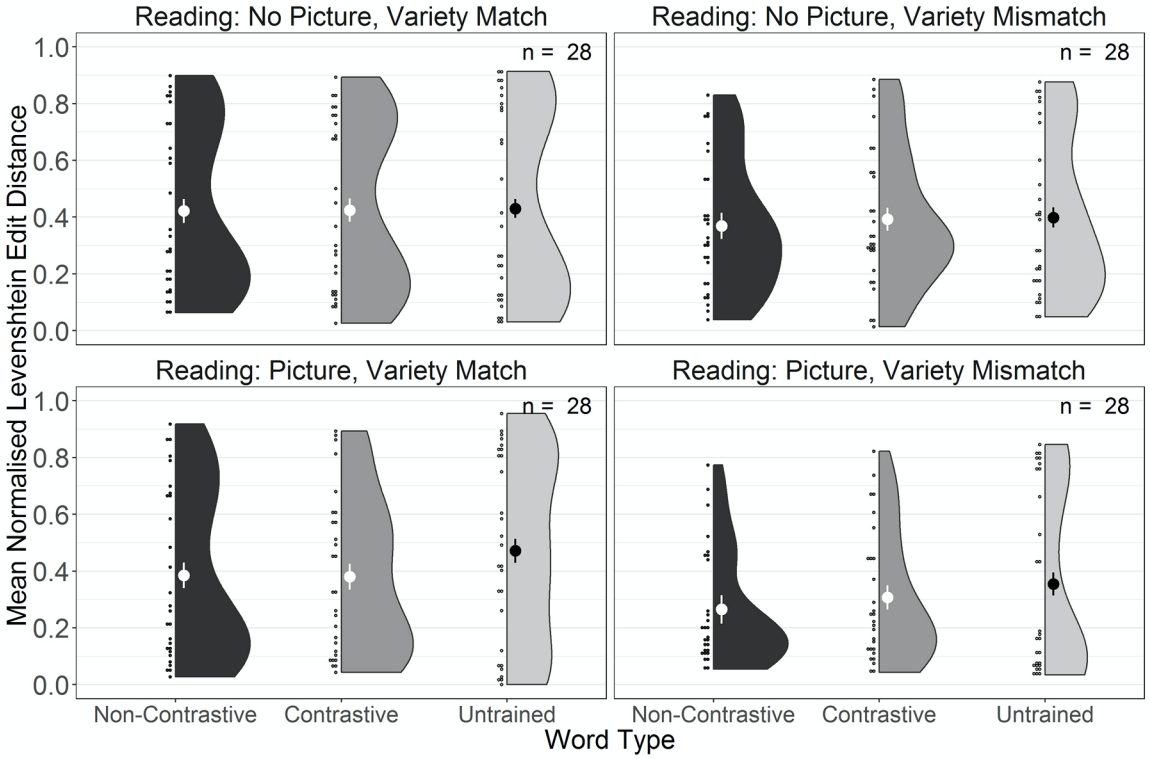

Figure 2: nLEDs for reading testing of trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 1. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

We performed a planned direct comparison of performance on untrained words only between the Variety Match and Variety Mismatch conditions. The model included fixed effects and interactions between the sum-coded Picture condition and Variety condition. We used the same criteria as in our main models for determining the random effects structure of the model. Here, this took the form of random zero-correlation intercepts and slopes of Picture condition and Variety condition and their interaction by items, and random intercepts by participants. This comparison showed no effect of Variety Mismatch (frequentist estimate: β = -0.05 [-0.13, 0.03], t = -1.21, p = .228; Bayesian Estimate: β = -0.09 [-0.26, 0.07]), thus failing to obtain conclusive evidence for a detrimental effect of dialect exposure on phonological decoding skills.

Discussion

In this experiment, participants learned to read 30 words of an artificial language using an artificial script. In the Variety Match condition, words presented during reading training were identical to words presented during exposure while in the Variety Mismatch condition half of the words varied between exposure and literacy acquisition mimicking dialect exposure. Half of the participants in each Variety condition saw pictures when

hearing and reading the words enabling them to develop semantic representations while the other half did not. Reading performance improved significantly over the course of training in both Variety conditions although the gains were steeper in the Variety Mismatch condition with pictures. We had predicted that performance would be worse for contrastive compared to non-contrastive words in the Variety Mismatch condition. While the results confirmed this trend, the contrastive deficit only reached significance during training in the No Picture condition, thus replicating findings from the reading experiment and the connectionist simulation of AAE exposure by Brown et al. (2015). Recall that in that simulation the contrastive deficit arose solely from similarity between the phonological representations of the AAE and MAE variants and not from competition between word forms associated with the same meaning. Our experiment was not able to unequivocally establish whether a contrastive deficit persist when meanings were provided by pictures as we only observed it in the No Picture condition during training but not reliably during testing.

We only observed a word familiarity benefit in the Variety Match condition with pictures. In natural languages, faster reading of high-frequency, familiar words compared to low-frequency words or non-words indicates the strength of the direct lexical route (Adelman et al., 2014; Caravolas, 2018). This lexicality effect is either due to more efficient, larger-grained processing of more familiar orthographic forms or the result of tighter links to word meanings in familiar words. In contrast, unfamiliar words require serial decoding of graphemes. In this experiment, links to word meanings could only be established in Picture conditions. In No-Picture conditions, benefits for trained words could arise either through greater acquired decoding efficiency or through partial decoding, e.g. when seeing the artificial equivalent of *BLEKUS*, participants may first decode *B* as /b/ and then *L* as /l/, at which point the phonological form /blekus/ (or the contrastive variant /blixus/ in the Variety Mismatch condition) may be recognised. The fact that the word familiarity benefit occurred only in picture conditions (reliably in the Variety Match and marginally in the Variety Mismatch condition) indcates that lexicality benefits arose only when access to phonological forms could be mediated by meanings. The absence of word familiarity effects in the No Picture conditions suggest that neither more efficient decoding strategies nor word recognition after partial decoding had a chance to emerge.

Our main question was whether exposure to competing dialect variants would affect learners' emerging phonological decoding skills. To answer this question, we compared reading performance for untrained words between the Variety Match and Mismatch conditions. If dialect exposure hinders reading skills in general, as suspected by the Head Teacher mentioned in our introductory paragraph, we would expect poorer performance with untrained words in the Variety Mismatch condition. Instead, we observed no difference to the Variety Match condition, although Bayesian estimates of the strength of evidence for the null hypothesis indicated that there was insufficient evidence for a null effect. We therefore can neither confirm nor exclude the possibility that dialect exposure impairs decoding skills.

As this experiment was a conceptual replication of the simulation of learning to read in Brown et al. (2015), it is not clear how well the findings of no difference between the Variety Match and Mismatch conditions generalise in the absence of spelling training. Moreover, in this experiment, learning grapheme-phoneme mappings was made difficult by the inconsistent orthography designed to mimic an orthography like English. Recall that we implemented two conditional rules according to which grapheme-phoneme and phoneme-grapheme mappings changed depending on context. These complex conditional rules likely further discouraged discovery and use of grapheme-phoneme conversion. To promote learning of such rules and to encourage phonologically mediated reading, we included spelling into Experiment 2. Participants were trained with an entirely consistent orthography in Experiment 2a and with the same inconsistent orthography in Experiment 2b, to examine whether effects of dialect exposure are similar for different levels of orthographic consistency.

Experiment 2: Effect of variety mismatch on learning to read and spell.

The aim of Experiment 2 was to provide more ecologically valid literacy training conditions by examining how exposure to variety mismatch affects learning to read and to spell a consistent (Experiment 2a) and an inconsistent (Experiment 2b) orthography.

Experiment 2a: Consistent orthography.

Method

Participants. One hundred and twelve participants (aged 20 – 65, aged M = 36.73, SD = 10.67, with 40 self-reported as female, 71 self-reported as male, and 1 self-reported as other) were recruited from Amazon's Mechanical Turk crowdsourcing platform and took part in the study for $7.50. Participants' mean English proficiency on a 1-5 Likert scale was 4.90 (SD = 0.40, range 2 - 5). Only eight participants rated their English proficiency as below 5. Eighty-seven participants reported knowing only English while 25 participants also knew Spanish (listed 12 times), French (listed 6 times), Hindi (listed 4 times) and eight other languages (listed a total of 11 times). Only one participant was familiar with a logographic script. Another two participants were tested and excluded based on the criteria described for Experiment 1.

Materials. We used the same set of graphemes, phonemes and words as in Experiment 1. In contrast to Experiment 1, we adopted only one-to-one mappings between graphemes and phonemes resulting in an entirely consistent orthography.

Procedure. The procedure was identical to Experiment 1 aside from the following deviation: During training, each ten-word block was presented once for reading and once for spelling (see Table 1). During spelling training participants heard a word and had to type it by clicking graphemes using an on-screen keyboard. Participants in the Picture condition always saw the picture of the associated referent when hearing the word. Once participants

had pressed the on-screen "Enter" key the correct spelling appeared below their own spelling for purposes of feedback. The feedback screen was cleared after 1.5 to 3.0 sec to prevent participants from taking notes or obtaining screenshots (the exact presentation time of the feedback was determined dynamically based on the word length, with a duration of 500ms per letter so that, for example, the correct spelling of a 4-letter-word would be presented for 2 sec). The overall amount of exposure to each item, combining presentations for reading and spelling, was identical to Experiment 1. In the testing phase, participants were presented with all thirty training words and an additional twelve untrained words in randomised order for reading and for spelling. Order of reading and spelling tasks was counterbalanced across participants but was kept constant across all phases within participants resulting in pseudo-random assignment of participants to 16 conditions comprising a crossing of Variety condition, Picture condition, speaker voice and task order. The mean completion time was 63.88 minutes (*SD* = 23.00).

Results

Coding. We used the same coding scheme for reading responses as in Experiment 1. The ICC between coders was F(111.00, 21.60) = 2754.73, $p < .001$, ICC = 0.999 [95% CI = 0.998; 1.000]. The 95% confidence interval around the parameter estimate indicates that the ICC falls above the bound of .90, which suggests excellent reliability across coders (Koo & Li, 2016). Spelling responses were analysed by computing length-normalised Levenshtein Edit Distances between response and target sequences of graphemes.

Model Fitting. Model fitting was similar to Experiment 1, with the exception of the inclusion of a sum-coded fixed effect of Task (reading vs. spelling) and of random slopes of Task. Additionally, since the training phase contained three training blocks per task, the training models were changed to include only an orthogonal linear (and not quadratic) time term as a fixed and random effect, to avoid overfitting change over time based on only 3 time points. Word Type was nested within the combination of Variety, Picture and Task conditions. We used maximal random effect structure comprising random intercepts and slopes of all fixed effects by participants and items, with zero-correlation between intercepts and slopes where appropriate to avoid non-convergence. The Bayesian mixed effects models used the same priors as in Experiment 1, with the addition of informative, *Normal* (0, 0.2) priors on the fixed effect of Task and any interactions of other terms with this factor.

Training. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 4. The results showed a main effect of Block, indicating an overall improvement of performance as training progressed, as well as a main effect of Task demonstrating better performance for reading than for spelling. Crucially, as indicated by the effect of Word Type, we found that reading, but not spelling, of contrastive words was significantly impaired in the Variety Mismatch conditions with and without pictures. In the Picture condition, the effect of Word Type in reading in the Variety Mismatch condition interacted with Block reflecting the fact

that impaired performance for contrastive words started to manifest itself gradually over the course of training (see Figures 3 and 4).

Table 4: Parameter estimates for the models fitted to nLEDs from the training phase in Experiment 2a. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Block (B) = 1-3, Variety Condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture Condition (PC) = picture vs. no picture (P vs. NP), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.66	0.04	[0.59, 0.73]	18.67	< .001	0.00	0.07	[-0.15, 0.14]
Block	**-8.92**	**0.77**	**[-10.44, -7.40]**	**-11.51**	**< .001**	**-0.37**	**0.03**	**[-0.44, -0.31]**
Task	**-0.04**	**0.01**	**[-0.06, -0.03]**	**-5.16**	**< .001**	**-0.09**	**0.02**	**[-0.12, -0.05]**
Picture Condition	0.05	0.03	[-0.02, 0.11]	1.33	.187	0.08	0.07	[-0.05, 0.21]
Variety Condition	-0.05	0.03	[-0.12, 0.02]	-1.43	.154	-0.09	0.07	[-0.21, 0.04]
B × T	-0.39	0.39	[-1.15, 0.38]	-0.99	.323	-0.02	0.02	[-0.05, 0.02]
B × PC	-0.27	0.77	[-1.79, 1.25]	-0.35	.725	-0.01	0.03	[-0.08, 0.05]
T × PC	0.00	0.01	[-0.02, 0.01]	-0.48	.635	-0.01	0.02	[-0.04, 0.02]
B × VC	0.90	0.77	[-0.62, 2.42]	1.16	.249	0.04	0.03	[-0.03, 0.10]
T × VC	-0.01	0.01	[-0.02, 0.01]	-1.19	.235	-0.02	0.02	[-0.05, 0.01]
PC × VC	-0.02	0.03	[-0.09, 0.05]	-0.55	.585	-0.03	0.07	[-0.16, 0.10]
B × T × PC	-0.18	0.39	[-0.94, 0.58]	-0.46	.645	-0.01	0.02	[-0.04, 0.02]
B × T × VC	0.01	0.39	[-0.75, 0.77]	0.02	.982	0.00	0.02	[-0.03, 0.03]
B × PC × VC	-1.43	0.77	[-2.95, 0.09]	-1.85	.067	-0.06	0.03	[-0.12, 0.01]
T × PC × VC	0.00	0.01	[-0.01, 0.02]	0.63	.533	0.01	0.02	[-0.02, 0.04]

B × T × PC × VC	-0.58	0.39	[-1.34, 0.18]	-1.49	.140	-0.02	0.02	[-0.06, 0.01]
R, NP, VMis: WT	**-0.04**	**0.01**	**[-0.07, -0.01]**	**-2.47**	**.015**	**-0.07**	**0.03**	**[-0.13, -0.01]**
S, NP, VMis: WT	0.00	0.01	[-0.03, 0.03]	0.27	.787	0.01	0.03	[-0.05, 0.07]
R, P, VMis: WT	**-0.05**	**0.01**	**[-0.08, -0.02]**	**-3.60**	**< .001**	**-0.10**	**0.03**	**[-0.16, -0.04]**
S, P, VMis: WT	-0.01	0.01	[-0.04, 0.02]	-0.67	.503	-0.02	0.03	[-0.07, 0.04]
R, NP, VMa: WT	-0.01	0.01	[-0.04, 0.01]	-0.99	.325	-0.02	0.03	[-0.08, 0.04]
S, NP, VMa: WT	-0.01	0.01	[-0.04, 0.02]	-0.55	.583	-0.01	0.03	[-0.07, 0.04]
R, P, VMa: WT	0.00	0.01	[-0.03, 0.03]	-0.02	.987	0.00	0.03	[-0.06, 0.06]
S, P, VMa: WT	-0.01	0.01	[-0.04, 0.02]	-0.49	.622	-0.01	0.03	[-0.07, 0.05]
B, R, NP, VMis: WT	-0.18	0.90	[-1.95, 1.59]	-0.20	.844	-0.01	0.04	[-0.09, 0.07]
B, S, NP, VMis: WT	-0.43	0.90	[-2.21, 1.34]	-0.48	.632	-0.02	0.04	[-0.09, 0.05]
B, R, P, VMis: WT	**-2.21**	**0.90**	**[-3.97, -0.45]**	**-2.46**	**.014**	**-0.09**	**0.04**	**[-0.16, -0.02]**
B, S, P, VMis: WT	-0.96	0.90	[-2.72, 0.80]	-1.07	.287	-0.04	0.04	[-0.11, 0.04]
B, R, NP, VMa: WT	-0.25	0.90	[-2.02, 1.51]	-0.28	.778	-0.01	0.04	[-0.08, 0.06]
B, S, NP, VMa: WT	0.55	0.90	[-1.22, 2.32]	0.61	.542	0.02	0.04	[-0.05, 0.10]
B, R, P, VMa: WT	-0.14	0.90	[-1.91, 1.63]	-0.15	.879	-0.01	0.04	[-0.08, 0.06]
B, S, P, VMa: WT	1.03	0.90	[-0.75, 2.80]	1.13	.257	0.04	0.04	[-0.03, 0.12]

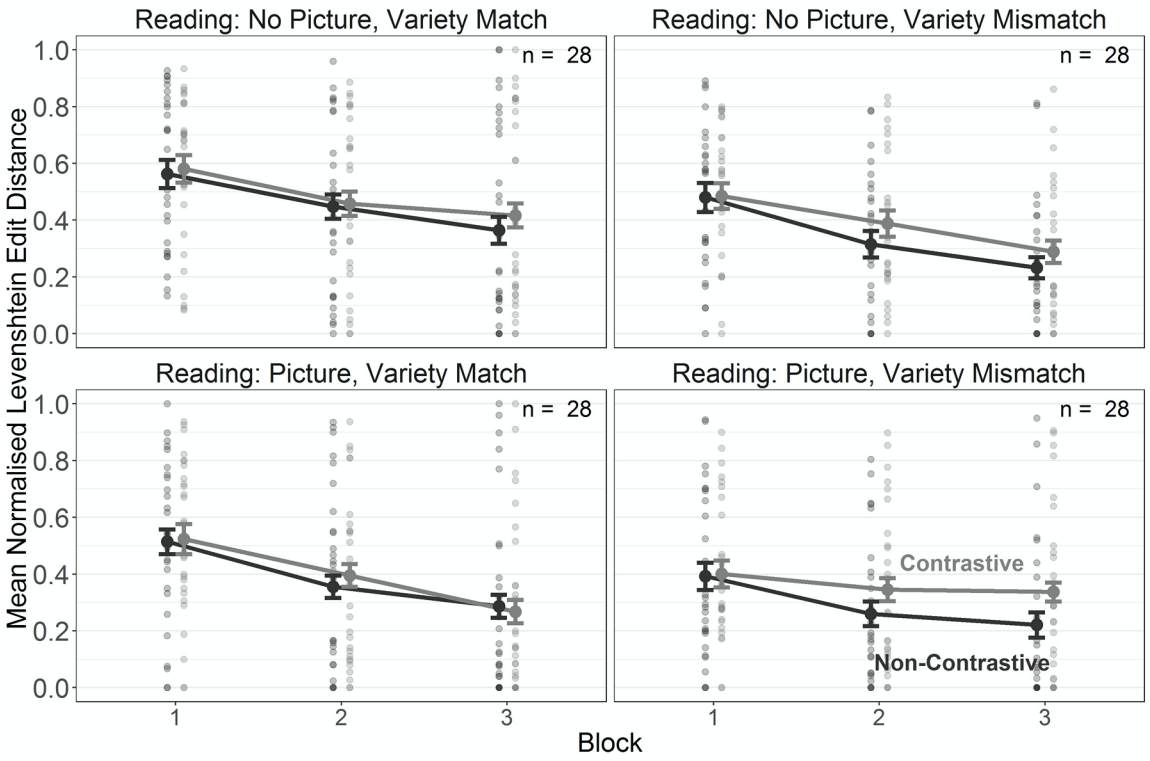

Figure 3: nLEDs for reading of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 2a. Error bars indicate ± 1 *SE* of the mean.

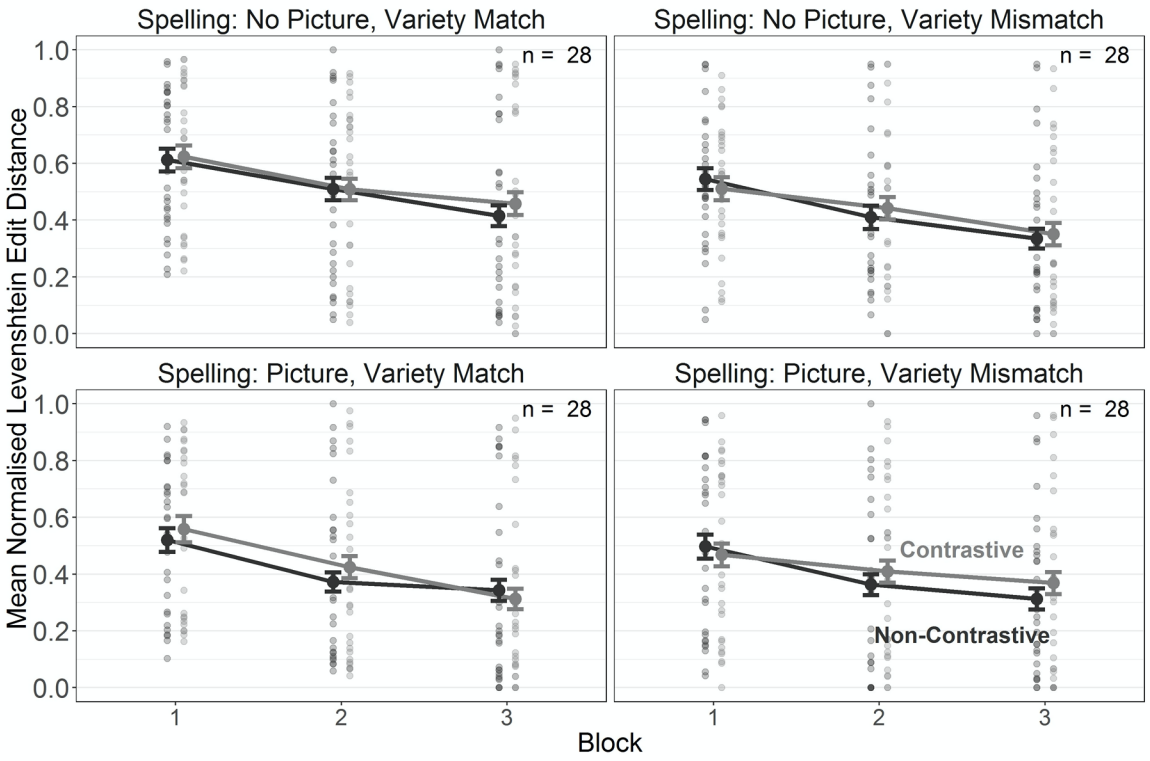

Figure 4: nLEDs for spelling of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 2a. Error bars indicate ± 1 *SE* of the mean.

Testing. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 5. The results confirmed the main effect of task observed during training which showed that performance was superior for reading compared to spelling. As during training, we found an effect of Word Type in the Variety Mismatch condition in reading but not in spelling, but only when pictures were present. The effect of Word Familiarity was significant in all conditions except for spelling in the Variety Mismatch condition with pictures, although Bayesian analyses failed to corroborate it for spelling in the Variety Match condition without pictures (see Figures 5 and 6).

Table 5: Parameter estimates for the models fitted to nLEDs from the testing phase in Experiment 2a. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture condition (PC) = picture vs. no picture (P vs. NP), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive, Word Familiarity (WF) = familiar vs. unfamiliar (novel)

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.52	0.04	[0.44, 0.60]	12.75	< .001	0.01	0.08	[-0.15, 0.16]
Task	**-0.04**	**0.01**	**[-0.06, -0.03]**	**-6.72**	**< .001**	**-0.08**	**0.01**	**[-0.11, -0.06]**
Picture Condition	0.05	0.04	[-0.03, 0.13]	1.14	.258	0.07	0.08	[-0.09, 0.21]
Variety Condition	-0.05	0.04	[-0.13, 0.03]	-1.15	.254	-0.09	0.07	[-0.24, 0.05]
T × PC	0.00	0.01	[-0.01, 0.01]	0.13	.894	0.00	0.01	[-0.02, 0.02]
T × VC	0.00	0.01	[-0.02, 0.01]	-0.72	.472	-0.01	0.01	[-0.03, 0.01]
PC × VC	-0.04	0.04	[-0.12, 0.04]	-1.03	.305	-0.07	0.07	[-0.21, 0.07]
T × PC × VC	0.00	0.01	[-0.01, 0.02]	0.76	.450	0.01	0.01	[-0.01, 0.03]
R, NP, VMis: WT	-0.02	0.02	[-0.06, 0.01]	-1.25	.216	-0.04	0.03	[-0.11, 0.02]
S, NP, VMis: WT	-0.01	0.02	[-0.04, 0.02]	-0.41	.685	-0.01	0.03	[-0.06, 0.05]
R, P, VMis: WT	**-0.05**	**0.02**	**[-0.09, -0.02]**	**-3.19**	**.002**	**-0.10**	**0.03**	**[-0.16, -0.04]**
S, P, VMis: WT	-0.01	0.02	[-0.04, 0.02]	-0.89	.377	-0.02	0.03	[-0.08, 0.03]
R, NP, VMa: WT	-0.02	0.02	[-0.05, 0.01]	-1.10	.274	-0.03	0.03	[-0.09, 0.03]
S, NP, VMa: WT	0.00	0.01	[-0.03, 0.02]	-0.15	.881	0.00	0.03	[-0.05, 0.05]
R, P, VMa: WT	0.00	0.02	[-0.04, 0.03]	-0.17	.869	0.00	0.03	[-0.06, 0.06]
S, P, VMa: WT	0.00	0.01	[-0.03, 0.02]	-0.10	.924	0.00	0.03	[-0.05, 0.05]
R, NP, VMis, WF	**0.03**	**0.01**	**[0.01, 0.06]**	**2.53**	**.013**	**0.06**	**0.02**	**[0.02, 0.11]**
S, NP, VMis, WF	**0.02**	**0.01**	**[0.00, 0.04]**	**2.21**	**.032**	**0.04**	**0.02**	**[0.01, 0.08]**
R, P, VMis, WF	**0.05**	**0.01**	**[0.03, 0.07]**	**4.07**	**< .001**	**0.09**	**0.02**	**[0.05, 0.14]**

S, P, VMis, WF	0.01	0.01	[-0.00, 0.03]	1.47	.149	0.03	0.02	[-0.01, 0.06]
R, NP, VMa, WF	**0.03**	**0.01**	**[0.01, 0.05]**	**2.58**	**.011**	**0.05**	**0.02**	**[0.01, 0.10]**
S, NP, VMa, WF	0.02	0.01	[0.00, 0.03]	2.03	.046	0.03	0.02	[-0.00, 0.07]
R, P, VMa, WF	**0.03**	**0.01**	**[0.00, 0.05]**	**2.35**	**.020**	**0.05**	**0.02**	**[0.01, 0.10]**
S, P, VMa, WF	**0.02**	**0.01**	**[0.00, 0.03]**	**2.48**	**.016**	**0.04**	**0.02**	**[0.00, 0.07]**

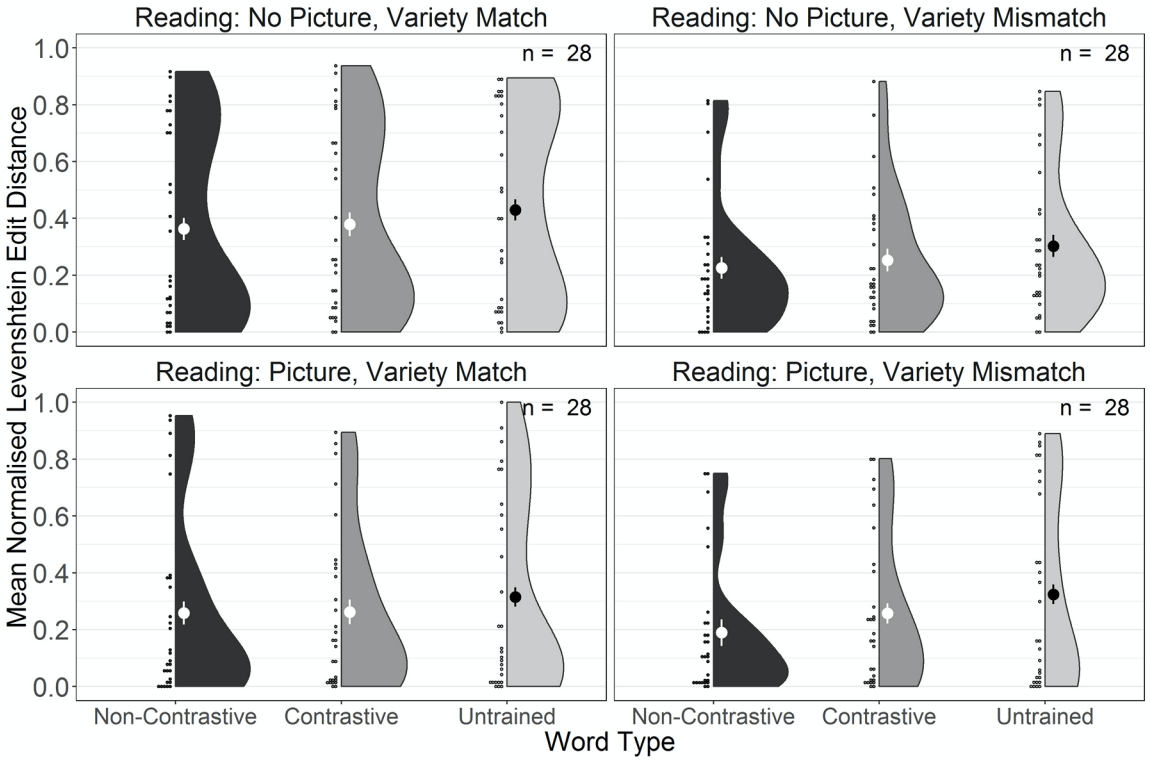

Figure 5: nLEDs for testing reading performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 2a. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

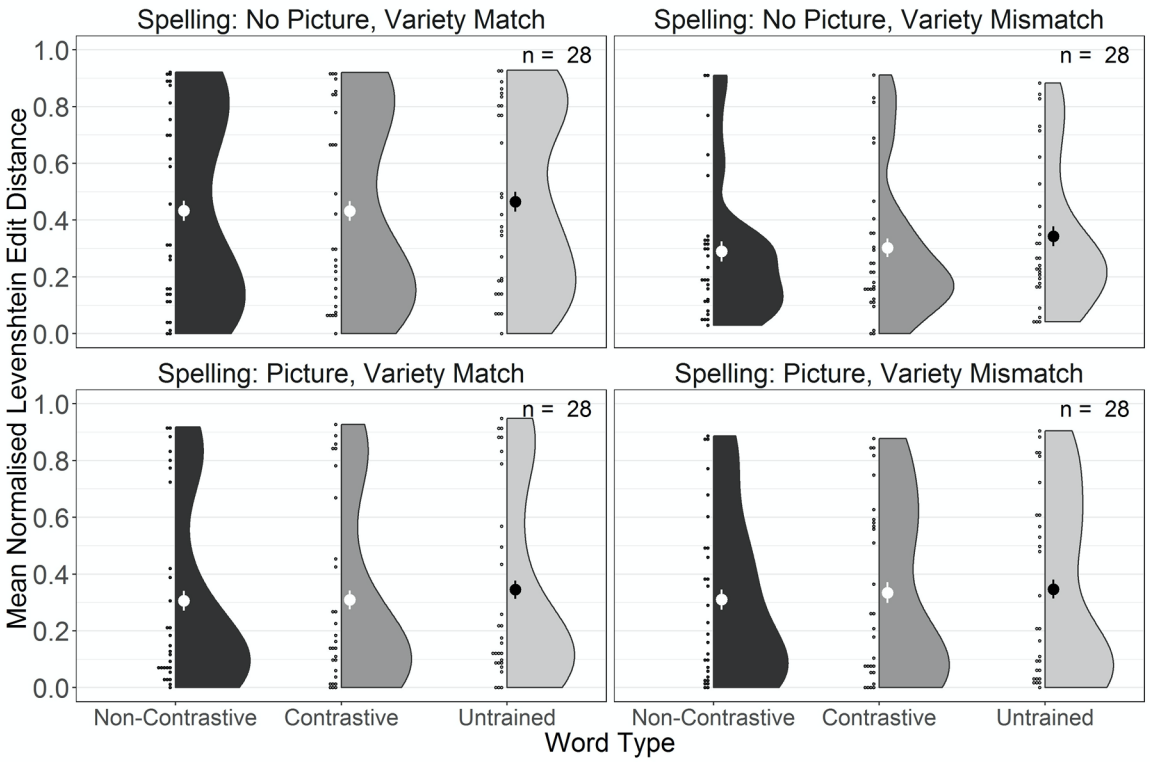

Figure 6: nLEDs for testing spelling performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 2a. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

As in Experiment 1, we performed a planned direct comparison of performance on untrained words between all Variety Match and Variety Mismatch conditions. The model included fixed effects and interactions between the sum-coded levels of Task condition, Picture condition and Variety condition. We used the same criteria as in our main models for determining the random effects structure. Here, this took the form of zero-correlation random intercepts and slopes of Task condition, Picture condition, Variety condition and their interaction by items, as well as random intercepts by participants. As in Experiment 1, the effect of Variety condition provided no evidence for a detrimental effect of a variety mismatch on reading and spelling of untrained words (frequentist estimate: β = -0.04 [-0.12, 0.04], t = -0.93, p = .353; Bayesian Estimate: β = -0.06 [-0.21, 0.07]).

Discussion

When spelling was introduced to literacy training in a consistent artificial orthography, the contrastive deficit emerged in reading when participants encountered a variety mismatch, which persisted into the testing session. However, unlike the contrastive reading deficit found by Brown et al. (2015) for children and neural networks exposed to both AAE and MAE here it was more persistent when meanings were provided by pictures.

Notably, no contrastive deficit emerged for spelling. This is because no competing orthographic representations for words in the exposure variety (i.e. the "dialect") existed and learners likely engaged in serial conversion of phonemes into graphemes. We had expected that introducing spelling would facilitate reliance on grapheme-phoneme decoding during reading, which should have attenuated the word familiarity effect. Yet the effect of word familiarity was significant in all reading conditions and even some of the spelling conditions. This is at odds with cross-linguistic findings of children learning to read (Caracolas, 2018), were lexicality effects were greater in the inconsistent orthography (English) compared to the consistent ones (Czech and Slovak). We suspect that the more consistent orthography may have encouraged more frequent partial decoding, i.e. decoding of just enough graphemes to access the memorised word form, which benefitted trained but not untrained items. In spelling, the word familiarity effect is somewhat puzzling but may reflect emerging representations of the overall graphemic Gestalt or even the motor routines required to type a word. Most relevant to the main question of the study, as in Experiment 1, similar reading and spelling performance with untrained words in the Variety Match and Mismatch conditions suggests that concurrent exposure to another variety did not seem to have any further detrimental effect on whatever decoding skills participants had acquired.

Experiment 2b: Inconsistent orthography.

Method

Participants. One hundred and twelve participants (aged 20 – 68, $M = 33.29$, $SD = 9.75$, with 38 self-reported as female and 74 self-reported as male) were recruited from Amazon's Mechanical Turk crowdsourcing platform and took part in the study for $7.50. Participants' mean proficiency in English on a 1-5 Likert scale was 4.77 ($SD = 0.57$, range 3 - 5). Only 18 participants rated their English proficiency as below 5. Seventy-one participants reported knowing only English while 41 participants also knew Spanish (listed 15 times), Hindi (listed 12 times), Tamil (listed 10 times) and 18 other languages (listed a total of 31 times). Only one participant was familiar with a logographic script. Another 3 participants were tested and excluded based on the criteria described for Experiment 1.

Materials. Graphemes, words and pictures were identical to the previous two experiments. We used the same inconsistent orthography as in Experiment 1.

Procedure. The procedure was identical to Experiment 2a. The mean completion time was 81.38 minutes ($SD = 42.82$).

Results

Coding. We used the same coding scheme for reading responses as in the previous experiments. The ICC between coders was $F(111.00, 84.86) = 2212.27$, $p < .001$, ICC = 0.999 [95% CI = 0.999; 0.999]. The 95% confidence interval around the parameter estimate indicates that the ICC falls above the bound of .90, which suggests excellent reliability

across coders (Koo & Li, 2016). Spelling responses were analysed by computing length-normalised Levenshtein Edit Distances between response and target sequences of graphemes.

Model Fitting. Frequentist and Bayesian analyses were conducted in the same way as for Experiment 2b. In the frequentist analyses, there were minor differences in the random effects structure compared to Experiment 2a due to differences in convergence: For the training phase, the maximal converging random effects structure included correlations between all by-participant terms. For the testing phase, correlations between all random effect terms had to be suppressed to avoid non-convergence.

Training. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 6. As in Experiment 2a, the main effect of Block indicated improvement in performance over the course of training and the main effect of Task confirmed that learning to spell was more difficult than learning to read. The only other significant effect was an interaction between Task and Picture condition. Pairwise contrasts based on the estimated marginal means of the training model were calculated using the *emmeans* R-package (Lenth, 2019), using Holm's sequential Bonferroni correction. These contrasts indicate that reading performance was better than writing performance in the picture condition only (Picture, Reading - Writing: ΔM = -0.12[-0.18, -0.05], t = -5.02, $p < .001$). All other contrasts were non-significant ($p > .05$). Unlike Experiment 1, we did not find any evidence for a contrastive deficit (see Figures 7 and 8).

Table 6: Parameter estimates for the models fitted to nLEDs from the training phase in Experiment 2b. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Block (B) = 1-3, Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture condition (PC) = picture vs. no picture (P vs. NP), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.80	0.03	[0.74, 0.87]	24.01	< .001	0.03	0.07	[-0.11, 0.17]
Block	**-6.44**	**0.68**	**[-7.78, -5.10]**	**-9.42**	**< .001**	**-0.28**	**0.03**	**[-0.34, -0.22]**
Task	**-0.03**	**0.01**	**[-0.05, -0.02]**	**-4.19**	**< .001**	**-0.07**	**0.02**	**[-0.11, -0.04]**
Picture Condition	-0.01	0.03	[-0.07, 0.05]	-0.18	.858	-0.01	0.06	[-0.13, 0.09]
Variety Condition	-0.04	0.03	[-0.10, 0.02]	-1.27	.207	-0.07	0.06	[-0.20, 0.06]

B × T	-0.28	0.39	[-1.05, 0.50]	-0.70	.487	-0.01	0.02	[-0.04, 0.02]
B × PC	-0.03	0.68	[-1.37, 1.31]	-0.04	.968	0.00	0.03	[-0.06, 0.06]
T × PC	**0.02**	**0.01**	**[0.01, 0.04]**	**2.87**	**.005**	**0.05**	**0.02**	**[0.01, 0.08]**
B × VC	-0.92	0.68	[-2.26, 0.42]	-1.35	.179	-0.04	0.03	[-0.10, 0.02]
T × VC	0.00	0.01	[-0.01, 0.02]	0.22	.824	0.00	0.02	[-0.03, 0.04]
PC × VC	0.01	0.03	[-0.05, 0.07]	0.45	.654	0.02	0.06	[-0.10, 0.14]
B × T × PC	0.66	0.39	[-0.11, 1.43]	1.67	.097	0.03	0.02	[-0.00, 0.06]
B × T × VC	-0.58	0.39	[-1.35, 0.20]	-1.46	.147	-0.03	0.02	[-0.06, 0.01]
B × PC × VC	0.40	0.68	[-0.94, 1.73]	0.58	.564	0.02	0.03	[-0.04, 0.07]
T × PC × VC	0.01	0.01	[-0.01, 0.02]	1.06	.293	0.02	0.02	[-0.02, 0.05]
B × T × PC × VC	-0.16	0.39	[-0.93, 0.61]	-0.40	.688	-0.01	0.02	[-0.04, 0.03]
R, NP, VMis: WT	-0.01	0.02	[-0.05, 0.03]	-0.51	.608	-0.02	0.04	[-0.09, 0.06]
S, NP, VMis: WT	-0.01	0.02	[-0.05, 0.02]	-0.71	.481	-0.02	0.04	[-0.10, 0.05]
R, P, VMis: WT	-0.01	0.02	[-0.05, 0.03]	-0.45	.655	-0.02	0.04	[-0.09, 0.06]
S, P, VMis: WT	0.01	0.02	[-0.02, 0.05]	0.80	.428	0.03	0.04	[-0.04, 0.10]
R, NP, VMa: WT	0.00	0.02	[-0.03, 0.04]	0.18	.861	0.01	0.04	[-0.07, 0.08]
S, NP, VMa: WT	-0.02	0.02	[-0.05, 0.02]	-0.87	.385	-0.03	0.04	[-0.10, 0.04]
R, P, VMa: WT	-0.02	0.02	[-0.06, 0.02]	-1.07	.288	-0.04	0.04	[-0.11, 0.03]
S, P, VMa: WT	-0.01	0.02	[-0.05, 0.02]	-0.81	.419	-0.03	0.04	[-0.10, 0.05]
B, R, NP, VMis: WT	-0.14	1.01	[-2.11, 1.84]	-0.14	.891	-0.01	0.04	[-0.09, 0.08]
B, S, NP, VMis: WT	0.48	1.02	[-1.52, 2.48]	0.47	.637	0.02	0.04	[-0.06, 0.11]
B, R, P, VMis: WT	0.29	1.00	[-1.68, 2.25]	0.29	.774	0.01	0.04	[-0.07, 0.10]

B, S, P, VMis: WT	-0.05	1.02	[-2.04, 1.94]	-0.05	.961	0.00	0.04	[-0.08, 0.08]
B, R, NP, VMa: WT	1.44	1.00	[-0.52, 3.40]	1.44	.150	0.06	0.04	[-0.02, 0.15]
B, S, NP, VMa: WT	0.64	1.01	[-1.34, 2.63]	0.63	.527	0.03	0.04	[-0.06, 0.11]
B, R, P, VMa: WT	0.90	1.01	[-1.08, 2.89]	0.89	.372	0.04	0.04	[-0.05, 0.12]
B, S, P, VMa: WT	0.50	1.02	[-1.50, 2.51]	0.49	.624	0.02	0.04	[-0.06, 0.10]

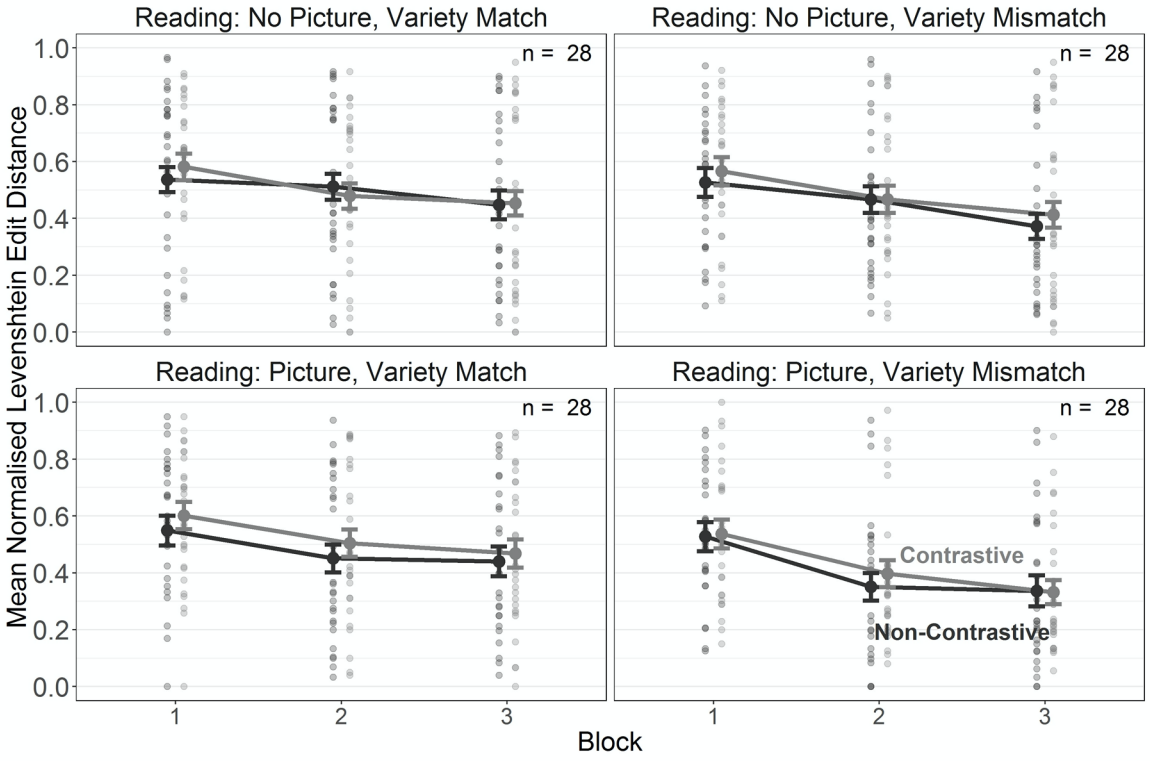

Figure 7: nLEDs for reading of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 2b. Error bars indicate ± 1 *SE* of the mean.

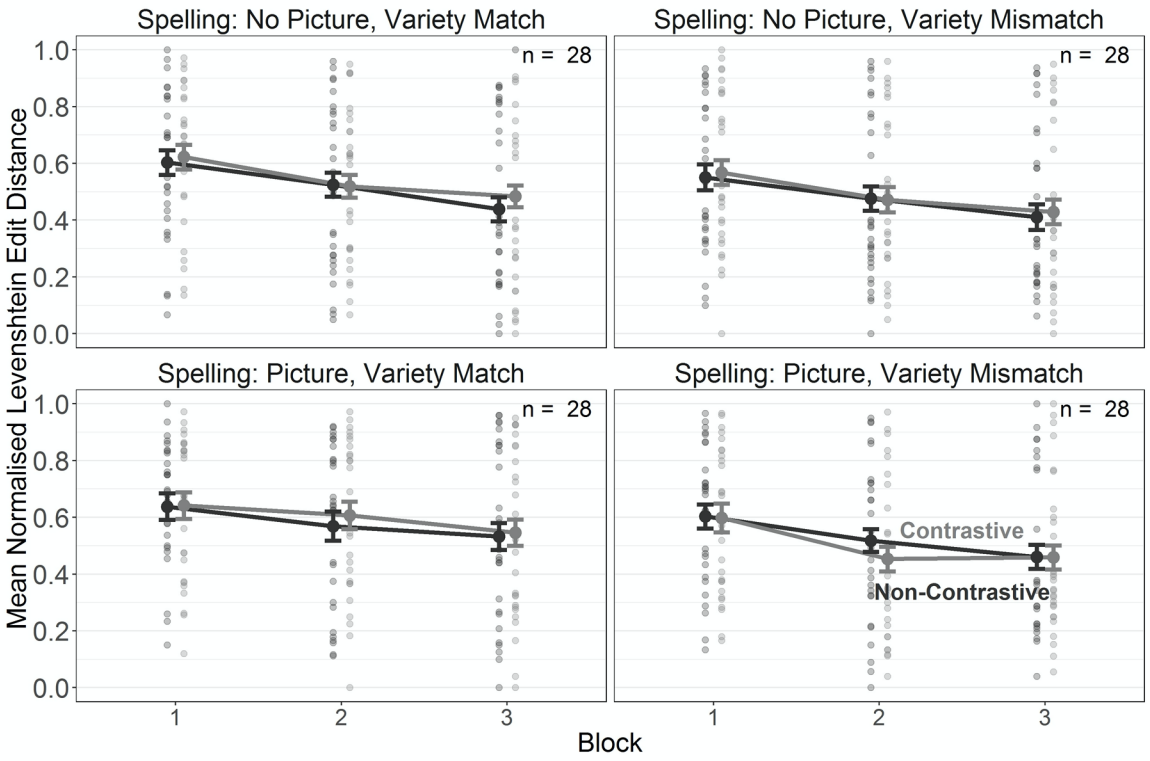

Figure 8: nLEDs for spelling of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 2b. Error bars indicate ± 1 *SE* of the mean.

Testing. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 7. The results confirmed the interaction between Task and Picture condition found already in the training data which suggests that reading performance was better than writing performance in the Picture condition only (Picture, Reading - Writing: ΔM = -0.11 [-0.16, -0.05], t = -5.40, $p < .001$). However, unlike Experiment 2b, there was no contrastive deficit and the effect of word familiarity appeared only in one condition, i.e. during reading in the Variety Match condition with pictures (see Figures 9 and 10).

Table 7: Parameter estimates for the models fitted to nLEDs from the testing phase in Experiment 2b. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Picture condition (PC) = picture vs. no picture (P vs. NP), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive, Word Familiarity (WF) = familiar vs. unfamiliar (novel)

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.69	0.04	[0.61, 0.77]	17.63	< .001	0.03	0.07	[-0.11, 0.18]
Task	**-0.03**	**0.01**	**[-0.04, -0.02]**	**-4.10**	**< .001**	**-0.06**	**0.01**	**[-0.09, -0.03]**
Picture Condition	0.00	0.04	[-0.07, 0.07]	0.06	.956	0.01	0.07	[-0.12, 0.14]
Variety Condition	-0.05	0.04	[-0.12, 0.02]	-1.31	.191	-0.08	0.07	[-0.23, 0.06]
T × PC	**0.02**	**0.01**	**[0.01, 0.04]**	**3.53**	**.001**	**0.05**	**0.01**	**[0.02, 0.07]**
T × VC	-0.01	0.01	[-0.02, 0.01]	-1.06	.292	-0.01	0.01	[-0.04, 0.01]
PC × VC	0.02	0.04	[-0.06, 0.09]	0.46	.647	0.03	0.07	[-0.10, 0.16]
T × PC × VC	0.00	0.01	[-0.02, 0.01]	-0.47	.641	-0.01	0.01	[-0.03, 0.02]
R, NP, VMis: WT	-0.01	0.02	[-0.05, 0.03]	-0.34	.731	-0.01	0.04	[-0.08, 0.07]
S, NP, VMis: WT	-0.02	0.02	[-0.06, 0.02]	-0.98	.327	-0.03	0.04	[-0.10, 0.04]
R, P, VMis: WT	-0.04	0.02	[-0.08, -0.00]	-2.06	.041	-0.07	0.04	[-0.15, 0.01]
S, P, VMis: WT	-0.01	0.02	[-0.05, 0.03]	-0.39	.695	-0.01	0.04	[-0.08, 0.06]
R, NP, VMa: WT	-0.02	0.02	[-0.06, 0.02]	-0.99	.322	-0.03	0.04	[-0.10, 0.04]
S, NP, VMa: WT	0.00	0.02	[-0.04, 0.04]	0.02	.986	0.01	0.04	[-0.06, 0.07]
R, P, VMa: WT	-0.03	0.02	[-0.07, 0.01]	-1.63	.105	-0.05	0.04	[-0.13, 0.02]
S, P, VMa: WT	0.00	0.02	[-0.04, 0.04]	-0.08	.935	0.00	0.03	[-0.06, 0.07]
R, NP, VMis, WF	0.01	0.02	[-0.03, 0.04]	0.44	.663	0.01	0.04	[-0.06, 0.08]
S, NP, VMis, WF	0.01	0.02	[-0.02, 0.04]	0.61	.541	0.02	0.03	[-0.04, 0.08]
R, P, VMis, WF	0.02	0.02	[-0.01, 0.05]	1.20	.231	0.04	0.04	[-0.03, 0.11]

S, P, VMis, WF	0.01	0.02	[-0.03, 0.04]	0.32	.747	0.01	0.03	[-0.05, 0.07]
R, NP, VMa, WF	0.02	0.02	[-0.01, 0.05]	1.08	.280	0.03	0.03	[-0.04, 0.10]
S, NP, VMa, WF	0.01	0.02	[-0.02, 0.04]	0.53	.598	0.01	0.03	[-0.04, 0.07]
R, P, VMa, WF	**0.04**	**0.02**	**[0.01, 0.08]**	**2.54**	**.012**	**0.08**	**0.03**	**[0.01, 0.15]**
S, P, VMa, WF	0.01	0.02	[-0.02, 0.04]	0.49	.624	0.01	0.03	[-0.04, 0.07]

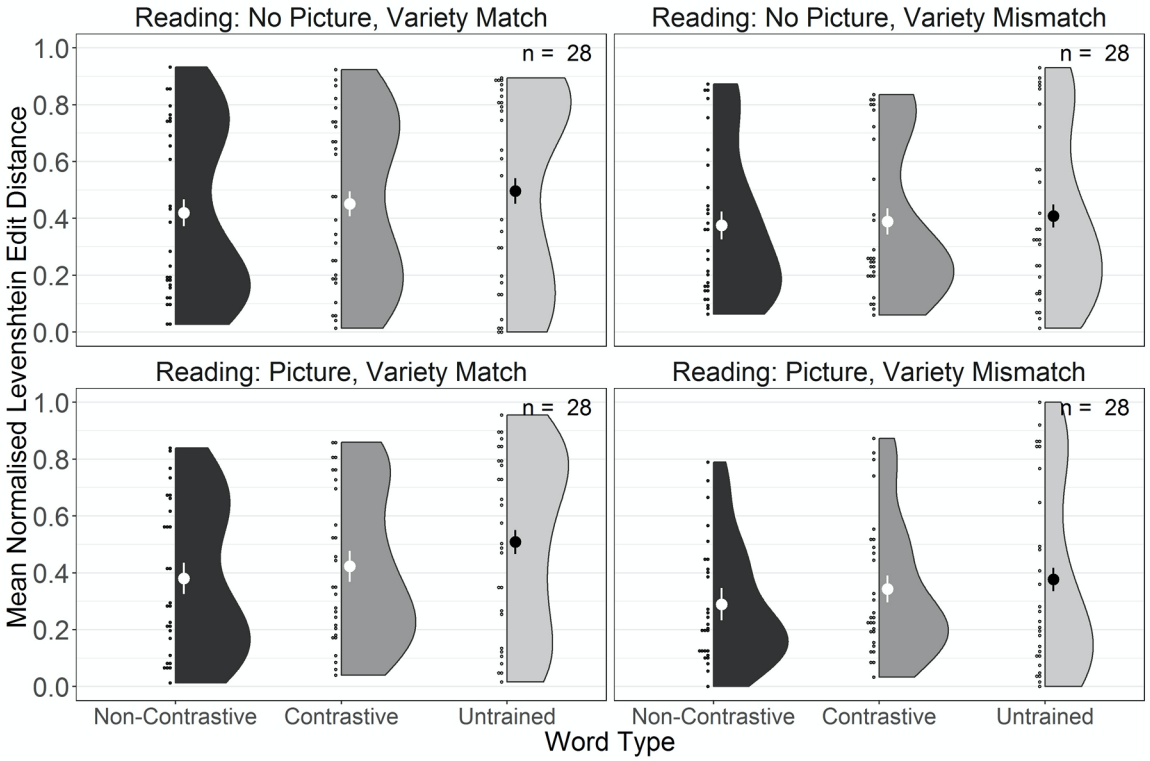

Figure 9: nLEDs for testing reading performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 2b. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

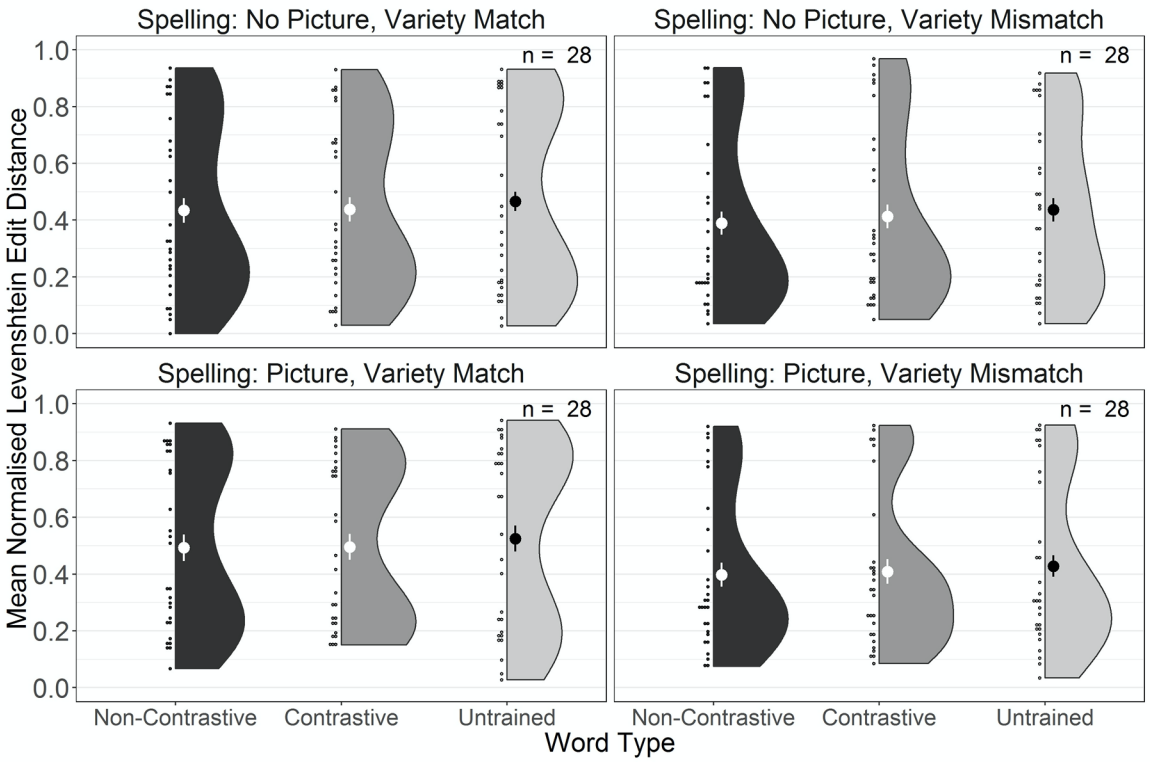

Figure 10: nLEDs for testing spelling performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 2b. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

The planned comparison of performance on untrained words only between all Variety Match and Variety Mismatch conditions used the same model structure as for Experiment 2a. There was no effect of Variety condition (frequentist estimate: β = -0.06 [-0.14, 0.02], t = -1.38, p = .171; Bayesian Estimate: β = -0.09 [-0.25, 0.07]), again suggesting that there was no evidence for a detrimental effect of exposure to a variety mismatch on reading and spelling of untrained words.

Discussion

When attempting to learn to read and to spell an inconsistent artificial orthography, participants showed improvement over the course of training. However, unlike under reading-only conditions in Experiment 1, where the contrastive deficit was found in some of the Variety Mismatch conditions, we found no contrastive deficit in this experiment. It is possible that learning conditional rules in reading and spelling rendered literacy acquisition too difficult to allow for the establishment of phonological representations that could have been placed into competition with each other, even when pictures provided meanings. Such an explanation is certainly in line with cross-linguistic studies of literacy acquisition in children, which show that at the early stages learning is more difficult for

inconsistent compared to more consistent orthographies (Seymour et al., 2003). In our experiment, were the artificial words were also novel, this may have hindered word learning; without more or less stable representations competition cannot occur. Again, as in Experiment 2a, there was no effect of variety mismatch on reading and spelling of untrained words suggesting that whatever weak decoding skills had been acquired remained unaffected by the presence of dialect variants words in the input.

Although this was not the main aim of this study, combining the first three experiments gives us the opportunity to explore whether orthographic consistency or spelling training are more conducive to literacy acquisition. Figure 11 shows a direct comparison of reading performance for all trained and untrained items during testing in the three experiments. To obtain statistical evidence for the comparison, we first fitted a linear mixed effect model with sum-coded fixed effects of Word Familiarity and treatment-coded fixed effects Experiment (1, 2a, 2b), and with a maximal random effects structure of random intercepts and slopes of Experiment by item and random intercepts and slopes of Word Familiarity by participants. Pairwise contrasts were then calculated for each experiment separately for trained vs. untrained words based on the estimated marginal means from the model using the *emmeans* R-package (Lenth, 2019). The results of these contrasts are provided in Table 8, where *p*-values are adjusted using Holm's sequential Bonferroni correction.

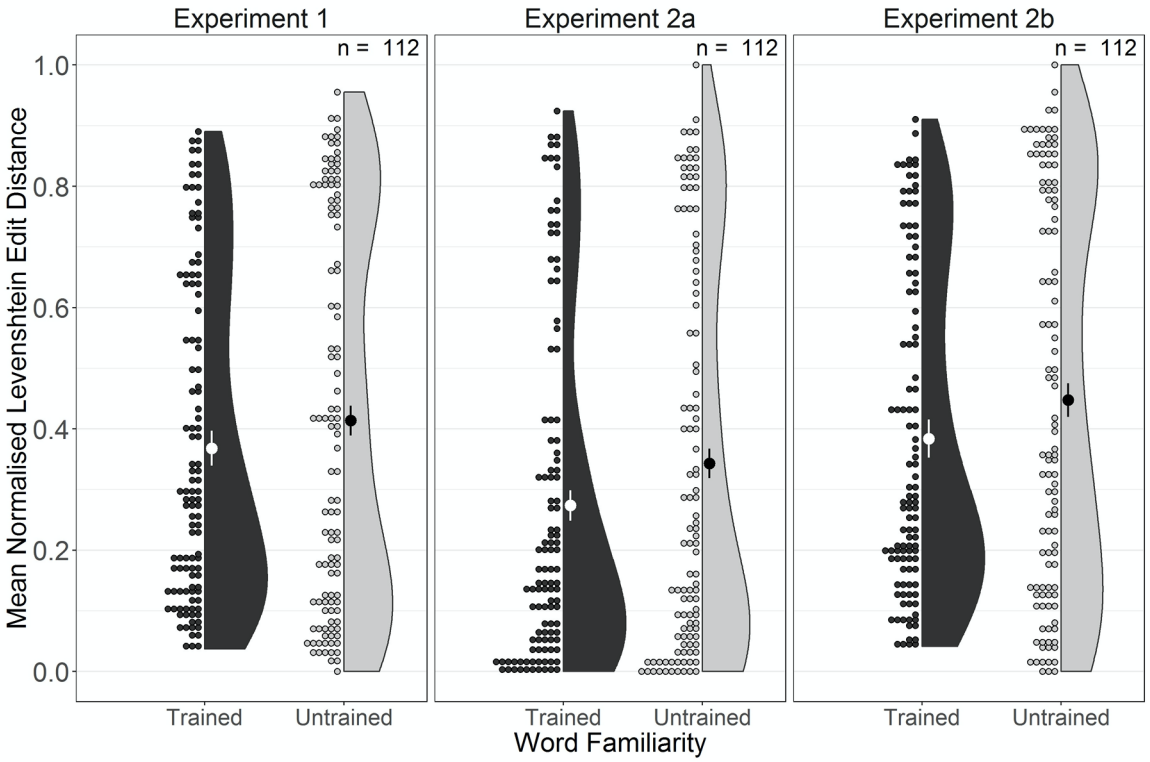

Figure 11: nLEDs for reading testing performance in trained and untrained words in Experiments 1, 2a, and 2b. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

These contrasts show that for trained words, performance was better in Experiment 2a compared to Experiment 1 and to Experiment 2b. All other differences were non-significant, suggesting that while introduction of spelling had no effect on overall reading outcomes, learning a consistent orthography led to measurable benefits, albeit only for trained words, regardless of whether spelling training was provided or not. Recall that the contrastive deficit also emerged most reliably with the consistent orthography suggesting that in this paradigm, phonological skill contributed to word learning, and competition between variants emerges only once learning has progressed to a stage at which access to phonological representations, either via (partial) decoding of the orthographic form or via semantic representations, is possible. However, the considerable variability in performance, evident in all figures, compellingly shows that participants differ tremendously in terms of their success at the early stages of this process. To create conditions that would allow for more reliable establishment of phonological representations, Experiment 3 repeated Experiment 2b with a longer training phase and a larger sample of learners, expecting to see a more reliable emergence of the contrastive deficit.

Table 8: Parameter estimates for pairwise contrasts between Experiments 1, 2a, and 2b as a function of Word Familiarity in the testing phase.

Contrast	Δ*M*	*SE*	95% Conf. I	*t*	*p*
Trained Words					
Experiment 1 - Experiment 2a	**0.09**	**0.02**	**[0.05, 0.13]**	**5.82**	**< .001**
Experiment 1 - Experiment 2b	-0.06	0.05	[-0.18, 0.05]	-1.34	.181
Experiment 2a - Experiment 2b	**-0.16**	**0.05**	**[-0.27, -0.04]**	**-3.22**	**.003**
Untrained Words					
Experiment 1 - Experiment 2a	0.06	0.03	[-0.00, 0.12]	2.42	.061
Experiment 1 - Experiment 2b	-0.07	0.05	[-0.20, 0.06]	-1.30	.193
Experiment 2a - Experiment 2b	-0.13	0.06	[-0.27, 0.00]	-2.32	.061

Experiment 3: Longer training of reading and spelling an inconsistent orthography.

Method

Participants. One hundred and sixty participants (aged 18 – 61, *M* = 32.48, *SD* = 9.67, with 89 self-reported as female, 70 self-reported as male, and 1 self-reported as other) were recruited from the crowdsourcing platform Prolific Academic and took part in the study for £9.00. All participants reported English as their native language and had a self-rated mean English proficiency on a 1-5 Likert scale of 4.86 (*SD* = 0.58, range 1 - 5). Participants reported no known mild cognitive impairments or dementia. Despite declaring English as their native language, 13 participants rated their English proficiency as below 5. Ninety-five participants reported knowing only English while 65 participants also knew French (listed 34 times), Spanish (listed 20 times), German (listed 12 times) and 26 other languages (listed a total of 50 times). Only eight participants were familiar with logographic scripts. An additional six participants were tested and excluded based on the exclusion criteria described for Experiment 1.

Materials. We used the same materials as in Experiments 1, 2a, and 2b, and the same inconsistent orthography as in Experiments 1 and 2b.

Procedure. The procedure deviated from Experiment 2a and 2b in that the training phase was doubled in length by adding another three ten-word reading and spelling blocks (with order of tasks counterbalanced across participants) resulting in a total of six training blocks for reading and spelling. All words were first partitioned into sets of ten for presentation in the first three reading and spelling blocks and then re-partitioned for presentation in the final three reading and spelling blocks, ensuring that each block contained five contrastive and five non-contrastive words. To provide ecologically valid conditions, semantic information was presented by depicting a concrete object with all words during exposure and reading training. The mean completion time was 98.14 minutes (*SD* = 91.20).

Results

Coding. We used the same coding scheme for reading responses as in the previous experiments. The ICC between coders was F(159.00, 159.86) = 635.90, $p < .001$, ICC = 0.997 [95% CI = 0.996; 0.998]. The 95% confidence interval around the parameter estimate indicates that the ICC falls above the bound of .90, which suggests excellent reliability across coders (Koo & Li, 2016). Spelling responses were analysed by computing length-normalised Levenshtein Edit Distances between response and target sequences of graphemes.

Model Fitting. We used a similar model structure to Experiments 2a and b, with the exclusion of the Picture condition factor. As in Experiments 1, 2a, and 2b, the fixed effects for training and testing were modelled by obtaining all main effects and interactions between all factors excluding Word Type, and nesting Word Type within each combination of factor levels of the Task and Variety conditions. Because this experiment, like Experiment 1, contained six blocks per task, we included the quadratic term for Block in the analyses of the training data to improve model fit. For the training data, the maximal converging random effects structure comprised zero-correlation random intercepts and slopes of Task, Variety condition, and their interaction by items, and random intercepts and slopes for the linear and quadratic time terms, Task, Word Type, and their interaction by participants, including all correlations between these terms. For the testing data, the random effects structure comprised random intercepts and slopes of Task, Variety condition, and their interaction by items, and random intercepts and slopes for Task, Word Type, and their interaction by participants, including all correlations between these terms for both by-participants and by-items random effects.

As with Experiments 1, 2a, and 2b, we also modelled the data using Bayesian mixed effects models with a full maximal random effects structure (i.e. without suppressing correlations between the by-items random effects in the training phase). These models used the same priors as in Experiments 2a and 2b, with the inclusion of a regularising, very weakly informative prior, *Normal*(0, 10), on the orthogonal quadratic time term, and excluding priors for Picture condition which was no longer in the model. We used these models to evaluate evidence in support of the null hypothesis for each parameter in the same way as in Experiments 1, 2a, and 2b.

Training. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 9. As in all previous experiments, we found a main effect of Block attesting performance improvement over the course of training. Similar to Experiment 1, the quadratic term also reached significance confirming non-linearity of the learning trajectory. We also confirmed the main effect of Task which indicates that reading performance exceeded spelling performance. The interaction between Block and Task suggests that while performance was similar across tasks at the outset, learning progressed more rapidly for reading than for spelling.

Table 9: Parameter estimates for the models fitted to nLEDs from the training phase in Experiment 3. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Block (B) = 1 - 6, Variety condition (VC) = variety match vs. variety mismatch (VMa vs. VMi), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive

	Frequentist Estimates	Bayesian Estimates

Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.73	0.03	[0.67, 0.78]	25.01	< .001	0.05	0.06	[-0.06, 0.15]
Block	**-17.99**	**0.86**	**[-19.68, -16.30]**	**-20.84**	**< .001**	**-0.63**	**0.03**	**[-0.68, -0.56]**
Block^2	**5.23**	**0.59**	**[4.07, 6.39]**	**8.84**	**< .001**	**0.18**	**0.02**	**[0.14, 0.22]**
Task	**-0.05**	**0.01**	**[-0.07, -0.03]**	**-5.50**	**< .001**	**-0.10**	**0.02**	**[-0.13, -0.06]**
Variety Condition	-0.03	0.03	[-0.08, 0.02]	-1.05	.296	-0.05	0.05	[-0.15, 0.04]
B × T	**-1.96**	**0.49**	**[-2.91, -1.00]**	**-4.01**	**< .001**	**-0.07**	**0.02**	**[-0.10, -0.03]**
B^2 × T	0.43	0.42	[-0.38, 1.25]	1.04	.298	0.01	0.01	[-0.01, 0.04]
B × VC	-1.70	0.86	[-3.39, -0.01]	-1.97	.051	-0.06	0.03	[-0.12, 0.00]
B^2 × VC	**1.71**	**0.59**	**[0.55, 2.87]**	**2.88**	**.004**	**0.06**	**0.02**	**[0.02, 0.10]**
T × VC	0.01	0.01	[-0.00, 0.03]	1.37	.171	0.02	0.01	[-0.01, 0.05]
B × T × VC	**1.00**	**0.49**	**[0.04, 1.95]**	**2.04**	**.043**	**0.03**	**0.02**	**[0.00, 0.07]**
B^2 × T × VC	-0.45	0.42	[-1.26, 0.37]	-1.08	.282	-0.02	0.01	[-0.04, 0.01]
R, VMis: WT	**-0.04**	**0.02**	**[-0.07, -0.01]**	**-2.61**	**.011**	**-0.07**	**0.03**	**[-0.13, -0.01]**
S, VMis: WT	0.00	0.02	[-0.03, 0.03]	-0.06	.952	0.00	0.03	[-0.06, 0.07]
R, VMa: WT	-0.02	0.02	[-0.05, 0.01]	-1.23	.222	-0.03	0.03	[-0.09, 0.03]
S, VMa: WT	0.00	0.02	[-0.03, 0.03]	-0.15	.882	0.00	0.03	[-0.06, 0.06]
B, R, VMis: WT	0.18	0.72	[-1.24, 1.60]	0.25	.802	0.01	0.02	[-0.04, 0.05]
B^2, R, VMis: WT	0.05	0.72	[-1.36, 1.46]	0.07	.945	0.01	0.02	[-0.03, 0.06]
B, S, VMis: WT	0.41	0.66	[-0.88, 1.71]	0.62	.534	0.00	0.02	[-0.04, 0.05]
B^2, S, VMis: WT	0.02	0.68	[-1.31, 1.36]	0.03	.973	-0.02	0.02	[-0.06, 0.03]
B, R, VMa: WT	0.01	0.72	[-1.41, 1.42]	0.01	.993	0.00	0.02	[-0.05, 0.05]

B^2, R, VMa: WT	-0.01	0.72	[-1.42, 1.39]	-0.02	.984	0.00	0.02	[-0.04, 0.05]
B, S, VMa: WT	-0.43	0.66	[-1.73, 0.86]	-0.66	.512	0.00	0.02	[-0.05, 0.04]
B^2, S, VMa: WT	-0.34	0.68	[-1.67, 1.00]	-0.49	.624	-0.01	0.02	[-0.06, 0.03]

With respect to the main questions of interest – the contrastive deficit and the effect of variety mismatch – we found evidence for a contrastive deficit for reading evidenced by the effect of Word Type in the Variety Mismatch condition. In addition, we observed an interaction between the quadratic term of Block and Variety condition and a three-way interaction between Block, Task and Variety condition, which suggest that performance levelled off somewhat faster in the Variety Match condition, especially for spelling, while further learning gains were made in the Variety Mismatch condition (see Figures 12 and 13).

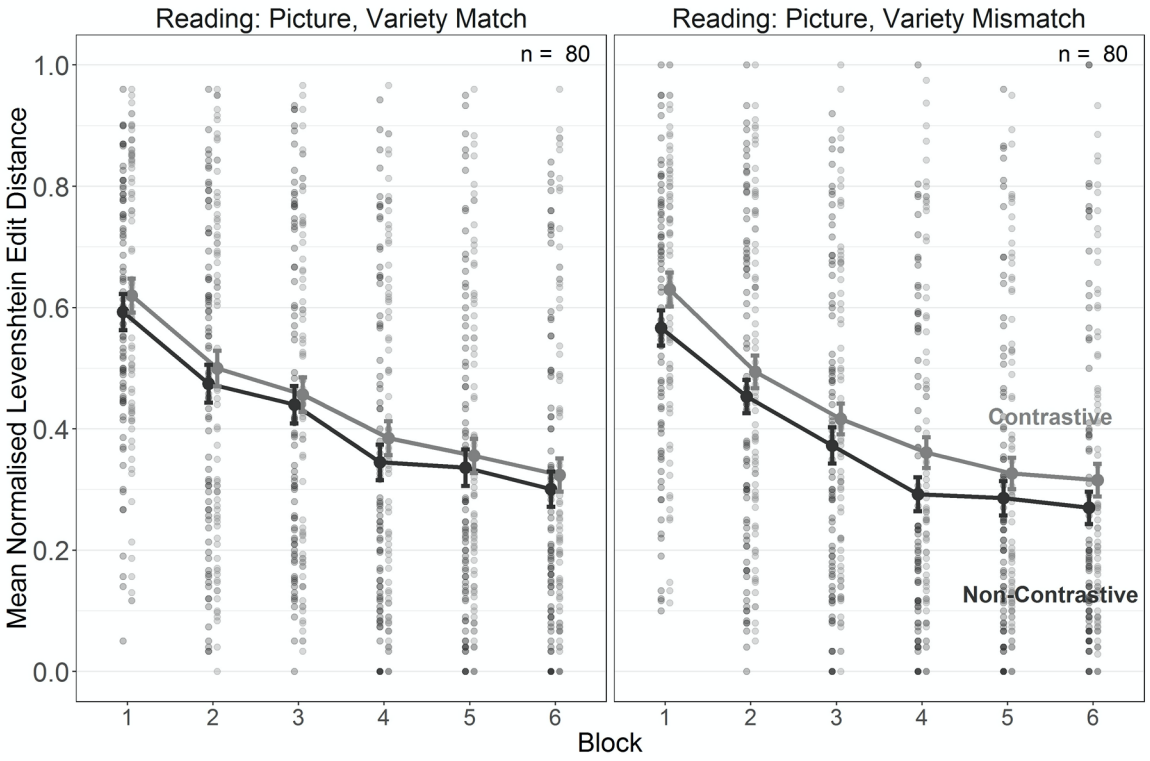

Figure 12: nLEDs for reading of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 3. Error bars indicate ± 1 *SE* of the mean.

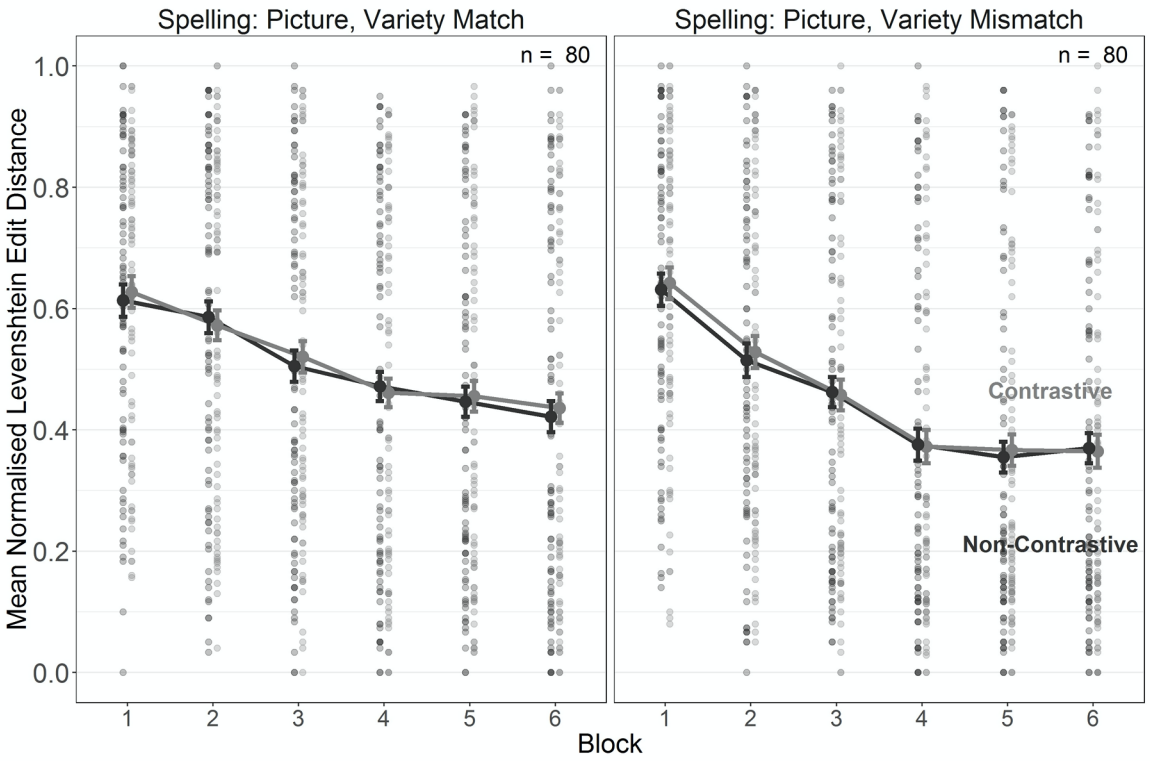

Figure 13: nLEDs for spelling of contrastive and non-contrastive words during 3 training blocks in the Variety Match and Variety Mismatch conditions in Experiment 3. Error bars indicate ± 1 *SE* of the mean.

Testing. Parameter estimates, confidence intervals (for the frequentist analysis) and credible intervals (for the Bayesian analysis) are presented in Table 10. As in the training data, there was a main effect of Task confirming superior performance for reading compared to spelling and an effect of Word Type, indicative of the contrastive deficit for reading in the Variety Mismatch condition. We also found that the effect of Word Familiarity was significant for reading in the Variety Match condition due to impaired performance for untrained compared to trained words in this condition. Crucially, the analysis yielded a main effect of Variety which showed that overall performance at test was superior in the Variety Mismatch condition (see Figures 14 and 15).

Table 10: Parameter estimates for the models fitted to nLEDs from the testing phase in Experiment 3. Bayesian analyses report standardised parameter estimates (i.e. the intercept [grand mean] is centred at 0). Values of 0 with a sign indicate the direction of the estimate before rounding. Variety condition (VC) = variety match vs. variety mismatch (VMa vs.

VMi), Task (T) = reading vs. spelling (R vs. S), Word Type (WT) = contrastive vs. non-contrastive, Word Familiarity (WF) = familiar vs. unfamiliar (novel)

	Frequentist Estimates					Bayesian Estimates		
Term	*Est.*	*SE*	95% Conf. I	*t*	*p*	*Est.*	*SE*	95% Cred. I
Intercept	0.58	0.03	[0.51, 0.64]	17.55	< .001	0.04	0.06	[-0.08, 0.17]
Task	**-0.05**	**0.01**	**[-0.06, -0.03]**	**-5.75**	**< .001**	**-0.09**	**0.02**	**[-0.12, -0.06]**
Variety Condition	**-0.07**	**0.03**	**[-0.12, -0.01]**	**-2.15**	**.033**	**-0.11**	**0.05**	**[-0.22, -0.01]**
T × VC	0.01	0.01	[-0.00, 0.03]	1.72	.088	0.02	0.01	[-0.00, 0.05]
R, VMis: WT	**-0.05**	**0.02**	**[-0.08, -0.02]**	**-2.86**	**.006**	**-0.09**	**0.03**	**[-0.16, -0.03]**
S, VMis: WT	0.00	0.02	[-0.04, 0.03]	-0.17	.865	0.00	0.03	[-0.06, 0.06]
R, VMa: WT	-0.03	0.02	[-0.06, 0.01]	-1.58	.121	-0.05	0.03	[-0.11, 0.02]
S, VMa: WT	-0.01	0.02	[-0.04, 0.03]	-0.31	.756	-0.01	0.03	[-0.07, 0.06]
R, VMis, WF	0.02	0.01	[-0.01, 0.04]	1.11	.270	0.03	0.03	[-0.02, 0.09]
S, VMis, WF	-0.01	0.01	[-0.03, 0.01]	-0.81	.423	-0.02	0.02	[-0.07, 0.03]
R, VMa, WF	**0.04**	**0.01**	**[0.01, 0.07]**	**2.82**	**.006**	**0.07**	**0.03**	**[0.02, 0.13]**
S, VMa, WF	0.00	0.01	[-0.02, 0.02]	-0.13	.897	0.00	0.02	[-0.05, 0.04]

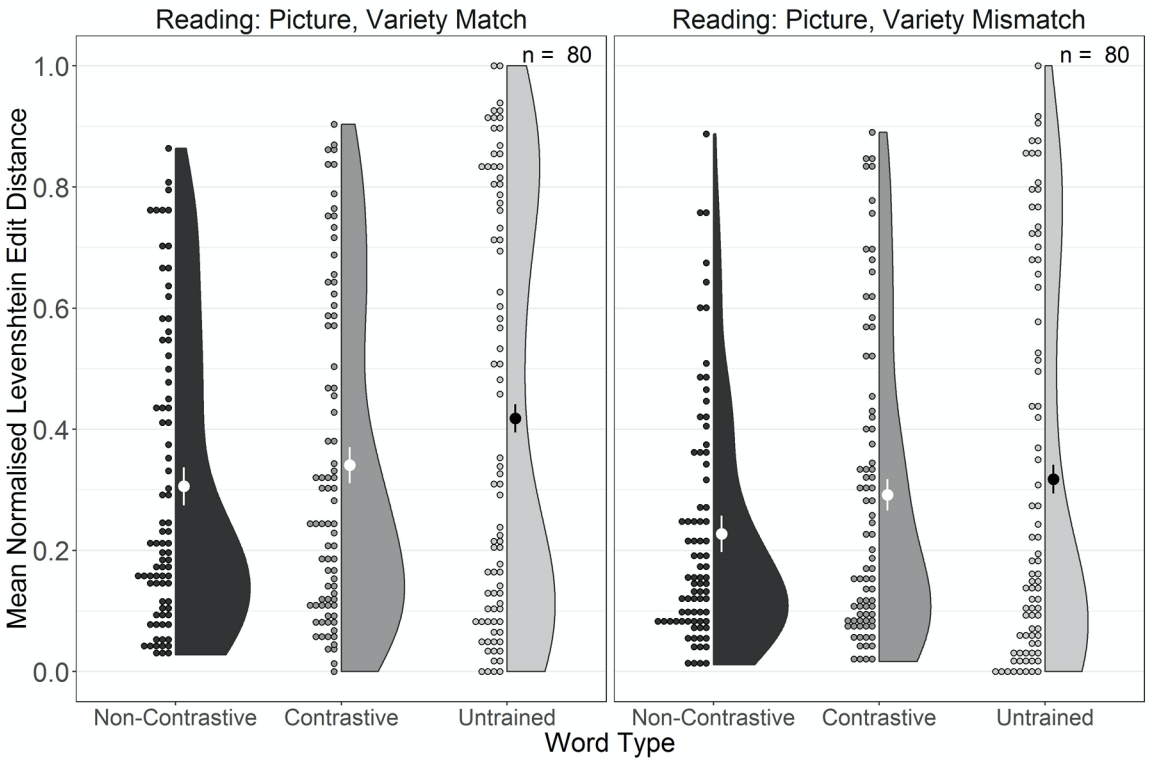

Figure 14: nLEDs for testing reading performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 3. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

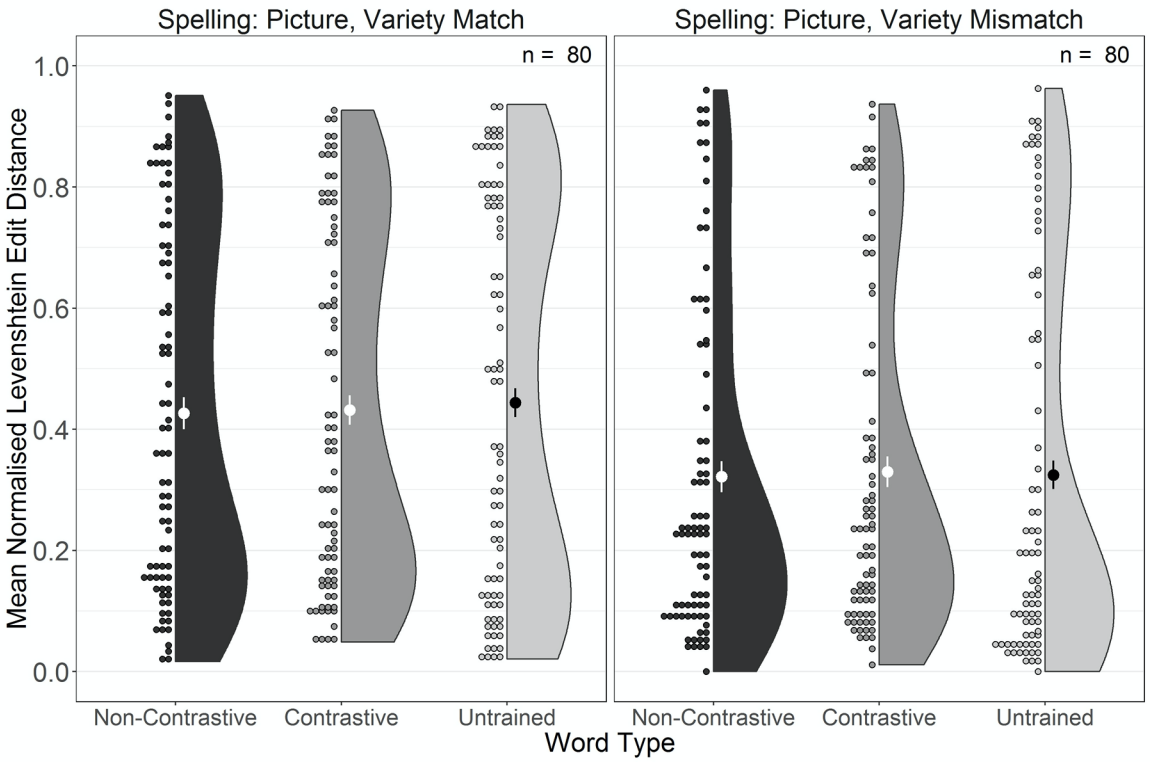

Figure 15: nLEDs for testing spelling performance for trained non-contrastive, trained contrastive and untrained words in the Variety Match and Variety Mismatch conditions in Experiment 3. Large dots and whiskers indicate means and ± 1 *SE* of the mean.

As in the previous experiments, we performed a planned comparison of performance between the Variety Match and Variety Mismatch conditions on untrained words only using the same model structure as in Experiments 2a and 2b. The frequentist model yielded a significant effect of Variety condition (β = -0.08 [-0.15, -0.01], t = -2.20, p = .029) with the Bayesian estimate suggesting sufficient evidence in favour of this effect (β = -0.14 [-0.26, -0.02]). This effect indicates that reading and spelling performance were superior in the Variety Mismatch condition.

Discussion

When a larger sample of participants was trained for a longer period in reading and spelling of an inconsistent artificial orthography with semantic information there was clear evidence for a contrastive deficit in reading, both in training as well as in testing. This indicates that when training is long enough for phonological representations to be established exposure of competing variants that are associated with the same meaning impairs reading. A contrastive deficit could not have arisen had participants exclusively relied on a phonologically mediated reading strategy that involved serial conversion of all graphemes into the associated phonemes. In contrast, no contrastive deficit emerged for

spelling because no dialect orthographic form had ever been presented, and because spelling could only be achieved through sequential conversion of individual phonemes into the associated graphemes.

At the same time, the word familiarity effect observed for reading in the Variety Match condition suggests that when no competing variants were encountered during literacy training participants seemed to rely more on direct access to the phonological forms of words, likely mediated by a word's meaning. In contrast, no phonological representations were available for untrained words making serial conversion of graphemes into phonemes necessary- a process that is presumably more error-prone than direct lexical access. The lack of an effect of Word Familiarity for reading in the Variety Mismatch condition suggests that having encountered many competing variants in the input discouraged a lexical strategy but rather encouraged grapheme-phoneme conversion, which was equally successful for trained and untrained words. As a result of more systematic use of phonological decoding, participants in the Variety Mismatch condition exhibited an overall benefit in their literacy skills, especially for untrained words.

General Discussion

In three experiments we investigated the effect of exposure to dialect variants of words on literacy learning. Employing an artificial language with an invented script allowed us to control for potential extra-linguistic confounds that are often associated with dialect exposure such as differences in quality of input, home literacy environment, cultural attitudes to literacy, educational provision or teacher expectation. Previous research (Brown et al., 2015) had shown that encountering a variety mismatch impairs processing of contrastive words, i.e. words with different variants across varieties (e.g. Scots /hoose/ vs. English /house/ or AAE /aks/ vs. MAE /ask/). What remained unclear was whether impaired performance with these contrastive words is also associated with a general deficit in decoding skills as measured by reading and spelling of novel, untrained words.

Our results confirmed and extended the finding of a contrastive deficit, which we replicated for reading training without semantic information in the Variety Mismatch conditions in Experiments 1 and 2a, where participants might not have noticed that similar, yet distinct variants were associated with the same lexical item. These conditions corresponded to the Brown et al. (2015) connectionist simulation where a contrastive deficit is akin to the processing of heterophonic homographs – words that are spelled the same but activate phonological competitors that are pronounced differently. However, when a consistent orthography (Experiment 2a) or longer training (Experiment 3) improved conditions for the establishment of phonological representations, the contrastive deficit appeared also when pictures enabled participants to access semantic representations, suggesting that a shared semantic representation further promotes competition between phonological variants, provided these are sufficiently stable. This finding is in line with interactive activation and competition models postulating inhibition from high-frequency competitors at the lexical layer, which can be reinforced via bidirectional connections

between lexical and semantic representations (Chen & Mirman, 2012), and suggests that both phonological and lexical competition contribute to a contrastive deficit in situations of dialect exposure.

One question that has not been addressed so far is whether the competition associated with the contrastive deficit in reading manifests itself in confusion between the two variants of a word or in increased non-specific errors when processing graphemic input. Our dependent variable, the Levenshtein Edit Distance, which provides the best comparison to the cross-entropy error computed for the neural network simulations of Brown et al. (2015), is not informative with respect to specific error types. In order to gain further insight into errors, we used automatic string comparison to code productions in the testing phase for all experiments with respect to whether dialect variants were produced in response to their standard contrastive counterpart (e.g. target: *kuble* – response: *xuble*), whether dialect variants were produced in response to another standard contrastive word (e.g. target: *skefi* – response: *xuble*), whether standard words were produced in response to another standard word (e.g. target: *skefi* – response: *kuble*), or whether any other non-substitution error was made (for summary graphs see Appendix E). Inspection of these response patterns shows a clear trend across experiments: while the mean percentage of correct responses to contrastive words did not differ between Variety Match (ranging from 27% to 44%) and Variety Mismatch (ranging from 28% to 45%) conditions, roughly three times more dialect variants were substituted for a standard contrastive counterpart (e.g. *kuble – xuble*) in Variety Mismatch (ranging from 5.5% to 8.2%) than Variety Match (ranging from 1.8% to 2.6%) conditions[8]. This trend suggests that the contrastive deficit, albeit small, is predominantly due to variant substitution rather than impaired overall reading skills.

We had hypothesised that introduction of spelling training should attenuate the contrastive deficit by facilitating phonologically mediated decoding (Taylor et al., 2017). Indeed, the fact that no contrastive deficit was found for spelling confirms that spelling itself did not rely on direct retrieval of orthographic forms but required conversion of phonemes into graphemes. In fact, variant substitution in response to contrastive words (e.g. *kuble – xuble*) did not occur at all for spelling even though spelling training did not prevent the emergence of such substitutions in reading, as described above. Moreover, as the joint analyses of Experiments 1 and 2 indicated, introducing spelling training did not lead to a significant improvement in overall literacy nor in phonologically mediated decoding, in contrast to studies demonstrating that invented, i.e. non-normative spelling facilitates reading by boosting phonemic awareness and by promoting a more analytical stance towards letter-sound correspondences (Caravolas et al., 2001; Ehri & Wilce, 2006; Ouellette & Sénéchal, 2008; Ouellette & Sénéchal, 2017; Ouellette et al., 2008). It is likely that adult learners, who already have mastered the alphabetic principle, do not experience an additional boost from spelling training as they prefer direct access to word forms during reading whenever possible – either after partial decoding of initial graphemes or via the

[8] Dialect errors in the Variety Match condition, where no dialect variants were never encountered, simply reflect the frequency with which these variants may occur if learners substitute or omit phonemes.

depicted word meaning or both. If conversion of individual phonemes into graphemes and vice versa is perceived as effortful and error-prone, adult participants may follow this route only when there is no alternative, as in spelling, for which performance was indeed consistently inferior to reading.

In the Picture conditions, we had included pictures not just during exposure but also during reading as a means of providing some semantic information to compensate for lack of sentential context or of accompanying pictures that often are found in children's books. It could be argued that whenever pictures were present alongside a word's orthographic form no decoding needed to take place at all as direct access of the phonological form via rote-memorisation of meaning-sound associations was possible. Under such conditions, dialect exposure should have no detrimental effects on the ability to decode novel words simply because no decoding skills would have been acquired, and reading of untrained words – the artificial-language analogy to non-word reading tests – should have presented considerable difficulty. Indeed, for the inconsistent orthography the familiarity effect was statistically significant when pictures were present, suggesting that a combination of a difficult-to-learn orthography with the availability of semantic information may have reduced the pressure to decode individual graphemes. However, it is unlikely that picture presentation during reading would have precluded the acquisition of decoding skills entirely because the number of times participants encountered each word and its meaning (five times in Experiments 1 and 2, ten times in Experiment 3) was probably insufficient to enable participants to reliably memorise sound-meaning associations for the entire set of 30 items, leaving them with having to decode, at least partially, those words they could not remember based on meaning. Reliance on partial decoding may then have been moderated by our experimental conditions: When pictures were provided, direct access from meaning to the sound form was possible, attenuating use of the decoding strategy. On the other hand, whenever a consistent orthography made decoding easier, as in Experiment 2a, word form access via meaning may have been discouraged, and partial decoding may have been encouraged so that a familiarity benefit appeared regardless of picture condition. Emergence of a word familiarity effect in some of the spelling conditions shows that a consistent orthography can facilitate access to orthographic representations, perhaps via implicit statistical learning of grapheme sequences, spatial locations of letters on the on-screen keyboard or associated motor routines that underpinned the keyboard-based spelling.

The crucial question of the present study was whether exposure to different variants of some of the training words in the Variety Mismatch conditions impaired decoding skills. In Experiments 1 and 2, Bayesian analyses indicated that there was insufficient evidence to answer this question. For the inconsistent orthography, this may simply have been a consequence of the overall difficulty of the task. But even for the consistent orthography, where learning was more successful, there was no evidence for a detrimental effect of variety mismatch. Moreover, when we increased our sample size to gain greater statistical power and extended the training phase (Experiment 3), we found a clear performance benefit in the Variety Mismatch condition. This benefit was significant for overall performance as well as for the untrained words separately, and provides clear evidence that

under conditions mimicking dialect exposure participants acquired superior decoding skills compared to conditions without dialect variation.

What might account for such a dialect benefit in artificial literacy acquisition? When discussing differential performance in reading and spelling we suggested that learners appear to select strategies based on perceived difficulty: We argue that greater linguistic variety may limit reliance on memory-based retrieval of phonological forms during reading and facilitate rule-based, phonologically mediated decoding, which, in turn, can lead to an overall improvement in decoding skills. This conclusion is also confirmed by the word familiarity effect in Experiment 3, which reached significance only in the Variety Match condition, suggesting that in the Mismatch condition, dialect exposure may have encouraged more reliance on phonological decoding to resolve the conflict between contrastive variants. Thus, counter to expectations formulated in the literature so far, our results suggest that when extra-linguistic confounds are controlled dialect exposure may in some situations even be beneficial for acquisition of phonological decoding skills.

Our experiments do not allow us to determine whether the observed dialect benefit requires explicit noticing of the competing variants for contrastive words or whether phonological decoding benefits simply from greater variability of word forms in the input. The idea that explicit noticing of dialect variants could benefit decoding skills is in agreement with the Linguistic Awareness/Flexibility Hypothesis of Terry and Scarborough (2011). While awareness of appropriate dialect usage can be seen as an indicator for general meta-linguistic knowledge, which is known to be beneficial for acquisition of phonological decoding skills, there is also evidence that directly boosting learners' dialect awareness can help literacy learning. For example, Johnson et al. (2017) demonstrated that an intervention that involved explicit teaching of dialect awareness to primary school children exposed to both NMAE and MAE resulted not only in more flexible and appropriate use of NMAE but also in better MAE literacy skills. Future research will have to investigate to what extent explicit awareness of contrastive words is required for a dialect benefit to occur during literacy learning.

Our finding of a dialect benefit in the artificial literacy paradigm comes with several caveats: First, learners in this study were adults who already had acquired literacy in one or more languages and were certainly familiar with the alphabetic principle. Their prior literacy competence may have endowed them with knowledge – implicit or explicit – of a variety of routes to access phonological and orthographic forms, and the ability to select strategically between them depending on input. Such choices may not be available to children who are just starting on the path to literacy using whatever principles are emphasised in their specific educational setting. Thus, caution is indicated when trying to generalise our findings to children until future research has examined whether dialect exposure has similar benefits in learners who are just beginning to acquire the different pathways to reading and spelling.

Secondly, the artificial conditions of our study differ from naturalistic literacy acquisition in several fundamental ways. For one, the goal of learning in the conditions in which no pictures were present was different from the typical goal of reading and spelling, which is to access and to convey meaning. Here, all that participants were asked to learn was the connection between print and sound, a limitation that was motivated by our attempt to replicate the findings from the Brown et al. (2015) connectionist simulations. Still, it may have shifted the emphasis on access to phonological and orthographic representations more than is appropriate in naturalistic literacy learning thereby affecting the learners' repertoire of mechanisms and strategies. We had tried to remedy this limitation by comparing these conditions with conditions in which pictorial information about the meaning was available at all times. However, unlike children, who typically know the meanings of the words they try to read and spell, in these conditions our participants learned the meaning of novel words at the same time as they learned to read and spell. This is more akin to acquisition of a second language in settings where learning is underpinned by print exposure, e.g. when adult speakers of English learn a Hebrew both from a teacher and a textbook – a more complex and potentially more effortful learning task than literacy learning in the native language. We had tried to mitigate against this additional burden by providing pictorial information about the meaning at all times, but it is still possible that this more difficult task may have proved taxing on attentional resources and thereby altered learning strategies. To be able to generalise from learning of artificial scripts to literacy acquisition in children future research may seek to study more ecologically valid conditions, for example the learning of novel scripts for familiar words or pre-training of word knowledge before literacy acquisition commences.

Thirdly, our experiments provided no cues, social or otherwise, for dialect use. All that participants encountered in the Variety Mismatch conditions was greater variability in terms of variants, whether associated with the same meaning or not. Yet dialect use is typically associated with specific regional, social and situational constraints. Brown et al. (2015), in their second simulation, showed that when dialect variants were cued by context nodes that coded variety (AAE vs. MAE) the contrastive deficit was attenuated. This shows that additional differentiating contextual information, provided consistently alongside phonologically similar contrastive variants, reduces competition. Despite the lack of social context, the present artificial language learning experiments are still of relevance as some evidence suggests that, unlike in bilingual language acquisition, the sociolinguistic competence required to contextualise dialect variation takes considerable time to build, as indicated by the slow developmental trajectory for dialect recognition (McCullough et al., 2019) and emergence of social attitudes towards dialects (Kinzler & DeJesus, 2013). One could construe the situation simulated in our experiments as one in which literacy acquisition precedes reliable acquisition of the sociolinguistic competence that governs dialect use. In future studies, we plan to provide contextual information alongside the different variants, which might reduce the difficulty with processing contrastive words. The intriguing question is in what ways such contextual information will affect the reliance on rote-memorisation vs. phonological decoding.

Finally, it should be noted that we observed considerable variability in performance in all experiments. A visual inspection of the figures indicates that the distributions of edit distances were bimodal in many conditions. Even though the lack of a normal distribution of this dependent variable does not preclude fitting the statistical models described above, as the residuals were normally distributed in all instances, it still points to the possibility qualitatively different mechanisms were employed by subgroups of our participants. This variability may in part reflect greater demographic diversity on crowdsourcing platforms compared to laboratory samples. We had refrained from selecting participants according to pre-specified demographic variables like SES because proxies for such variables (e.g. annual income) may have different validity in different cultural and economic contexts, and because of evidence that on crowdsourcing platforms responses to eligibility questions may not be reliable and consistent (Chandler & Paolacci, 2017). (Note in this context the curious discrepancy in some participants who were asked to self-select as native English speakers in Experiments 1 and 3 but rated their English proficiency as below-native or even elementary). Variability in performance may also reflect different solutions to the trade-off between minimising expended effort while maximising monetary gain, which may depend on whether participants use crowdsourcing platforms repeatedly as a source of income (El Maarry et al., 2018). In particular, the substantial duration of our experiments, in conjunction with the monetary reward, may have induced effort-minimising strategies beyond what would be expected in more naturalistic literacy acquisition contexts and in potentially better supervised laboratory studies. Although we tried to mitigate against outright cheating (e.g. note-taking) by placing time constraints on different tasks, we still have to accept that some participants may have expended too little effort for learning to occur. These shortcomings should at least in part be compensated for by our substantial sample sizes that exceed those typically used in laboratory experiments.

Conclusions

In naturalistic contexts, it is difficult to disentangle dialect exposure from other confounding factors that may affect literacy learning. The results from this artificial literacy learning study showed that while words with dialect variants are more difficult to read, their presence in the input can facilitate acquisition of phonological decoding skills as a means of reducing the arising competition. Because a phonologically mediated route to literacy acquisition has been shown to be essential in the early stages of learning to read and spell (Castles et al., 2018; Taylor et al., 2017) our results – if confirmed in further studies with children – raise the intriguing possibility that dialect exposure may, in fact, yield tangible benefits for literacy acquisition.

Context of the Research

This project has brought together two strands of experimental research that we have pursued in the past: the study of how cognitive representations of dialects in bidialectal speakers differ from representations of languages in bilinguals, and the study of how distributional features of the language input affect language learning. Inspired by the

applied question of whether dialect bans in schools are justified from the point of view of the underlying learning mechanisms, we extended the artificial language learning paradigm to the investigation of how input variability induced by dialect exposure might affect literacy acquisition. A major challenge was to scale up artificial language learning to larger numbers of participants via the use of crowd-sourcing platforms. To our knowledge, this is the first study to analyse large scale artificial language production data obtained from online participants. The strict controls afforded by artificial language and artificial script learning enabled us to replicate with human learners what neural network simulations had demonstrated before for natural language: that there is a small cost for processing words for which dialect variants exist. Our finding that this local cost does not necessarily impair acquisition of general decoding skills, at least in adult learners, will hopefully be of interest to researchers working on artificial language learning, on models of bidialectal lexical representation and on literacy acquisition as well as to educational practitioners. In the future, we will aim to extend this controlled approach to the study of how dialect exposure affects literacy acquisition in children.

References

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology*, *49*(7), 1348–1365. doi:10.1037/a0029839

Artiles, A. J., Kozleski, E. B., Osher, D., & Ortiz, A. (2010). Justifying and explaining disproportionality, 1968–2008: A Critique of Underlying Views of Culture. *Exceptional Children*, *76*(3), 279–299. doi:10.1177/001440291007600303

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

BBC News. (2013). Colley Lane school in Halesowen bans Black Country dialect. Retrieved from https://www.bbc.co.uk/news/uk-england-birmingham-24941692

Boersma, P., & Weenik, D. (2017). Praat: Doing phonetics by computer. Retrieved from http://praat.org/

Bolker, B., & Robinson, D. (n.d.). *Broom.mixed: Tidying methods for mixed models*. Retrieved from http://github.com/bbolker/broom.mixed

Bowers, J. S., & Bowers, P. N. (2017). Beyond phonics: The case for teaching children the logic of the English spelling system. *Educational Psychologist*, *52*(2), 124–141. doi:10.1080/00461520.2017.1288571

Bowers, J. S., & Bowers, P. N. (2018). Progress in reading instruction requires a better understanding of the English spelling system. *Current Directions in Psychological Science*, *27*(6), 407–412. doi:10.1177/0963721418773749

Brown, M. C., Sibley, D. E., Washington, J. A., Rogers, T. T., Edwards, J. R., MacDonald, M. C., & Seidenberg, M. S. (2015). Impact of dialect use on a basic component of learning to read. *Frontiers in Psychology*, *6*, 1–17. doi:10.3389/fpsyg.2015.00196

Bühler, J. C., Oertzen, T. von, McBride, C. A., Stoll, S., & Maurer, U. (2018). Influence of dialect use on early reading and spelling acquisition in German-speaking children in Grade 1. *Journal of Cognitive Psychology*, *30*(3), 336–360. doi:10.1080/20445911.2018.1444614

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in Psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1–25. doi:10.1177/2515245918823199

Caravolas, M. (2018). Growth of word and pseudoword reading efficiency in alphabetic orthographies: Impact of consistency. *Journal of Learning Disabilities*, *51*(5), 422-433. doi:10.1177/0022219417718197

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language*, *45*(4), 751-774. doi: 0.1006/jmla.2000.2785

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. doi:10.1177/1529100618772271

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500-508. doi: 10.1177/1948550617698203

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B*, *272*, 267–75. doi:10.1098/rspb.2004.2942

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school English in African American children and its relation to early reading achievement. *Child Development, 75*(5), 1340–1356.

Chen, Q., & Mirman, D. (2012). Competition and Cooperation Among Similar Representations: Toward a Unified Account of Facilitative and Inhibitory Effects of Lexical Neighbors. *Psychological Review*, *119*(2), 417–430. doi:10.1037/a0027175

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. doi:10.1037//0033-295x.108.1.204

Crystal, D. (2003). *The Cambridge Encyclopedia of the English language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, *100*(4), 907–919. doi:10.1037/a0012656

Dictionary of the Scots Language. (n.d.). Muckle. Retrieved from http://www.dsl.ac.uk/entry/snd/muckle

Donaldson, J. (1999). *The Gruffalo*. Pan Macmillan.

Donaldson, J. (2005). *The Gruffalo's Child*. Pan Macmillan.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393-447.

Ehri, L. C., & Wilce, L. S. (2006). Does learning to spell help beginners learn to read words? *Reading Research Quarterly*, *22*(1), 47–65. doi:10.2307/747720

Ellis, N., & Cataldo, S. (1990). The role of spelling in learning to read. *Language and Education*, *4*(1), 1-28. doi:10.1080/09500789009541270

El Maarry, K., Milland, K., & Balke, W.-T. (2018). A fair share of the work? The evolving ecosystem of crowd workers. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 145–152).

Forsythe, A., Street, N., & Helmy, M. (2017). Revisiting Rossion and Pourtois with new ratings for automated complexity, familiarity, beauty, and encounter. *Behavior Research Methods*, *49*(4), 1484–1493. doi:10.3758/s13428-016-0808-z

Fox, J., Venables, B., Damico, A., & Salverda, A. P. (2019). *English: Translate integers into english*. Retrieved from https://CRAN.R-project.org/package=english

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, M. Coltheart, & J. Marshall (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). London: Erlbaum.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Irr: Various coefficients of interrater reliability and agreement*. Retrieved from https://CRAN.R-project.org/package=irr

Gatlin, B., & Wanzek, J. (2015). Relations among children's use of dialect and literacy skills: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *58*, 1306–1318. doi:10.1044/2015

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 561–574. doi:10.1037/00961523.25.2.561

Harber, J. R. (1977). Influence of presentation dialect and orthographic form on reading performance of Black, inner-city children. *Educational Research Quarterly*, *2*(2), 9–16.

Harley, H. (2006). *English words: A linguistic introduction*. Oxford, UK: Blackwell Publishing.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720. doi:10.1037/0033-295X.111.3.662

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology*, *20*(2), 115–162. doi:10.1080/02643290242000871

Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2012). Cross-language activation of phonology in young bilingual readers. *Reading and Writing*, *25*(6), 1327–1343. doi:10.1007/s11145-011-9320-0

Johnston, P. A. J. (2007). Scottish English and Scots. In D. Britain (Ed.), *Language in the British Isles* (pp. 105–121). Cambridge, UK: Cambridge University Press.

Johnson, L., Terry, N. P., Connor, C. M., & Thomas-Tate, S. (2017). The effects of dialect awareness instruction on nonmainstream American English speakers. *Reading and Writing, 30*(9), 2009-2038.

Kimball, S., Mattis, P., & The GIMP Development Team. (1995). GNU Image Manipulation Program. Retrieved from https://www.gimp.org/

Kinzler, K. D., & DeJesus, J. M. (2013). Northern = smart and Southern = nice: The development of accent attitudes in the United States. *Quarterly Journal of Experimental Psychology*, *66*(6), 1146–1158. doi:10.1080/17470218.2012.731695

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. doi:10.1016/j.jcm.2016.02.012

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*, 178–206. doi:10.3758/s13423-016-1221-4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Labov, W. (1995). Can reading failure be reversed? A linguistic approach to the question. In V. L. Gadsden & D. A. Wagner (Eds.), *Literacy among African-American youth: Issues in learning, teaching, and schooling* (pp. 39–68). Cresskill, NJ: Hampton Press.

Lenth, R. (2019). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 8,* 707-710.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

Mazzoni, D., & Dannenberg, R. (2016). Audacity. Retrieved from http://audacityteam.org/

McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, *62*(1), 115–136. doi:10.1177/0023830917743277

Milde, B. (2011). Shapecatcher: Unicode character recognition. Retrieved from http: //shapecatcher.com/

Mirman, D. (2014). *Growth curve analysis and visualization Using R* (p. 168). Boca Ranton, FL.: Chapman; Hall/CRC Press.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN. R-project.org/package=here

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Linguistics and Language Compass*, *10*(11), 591–613. doi:10.1111/lnc3.12207

Ouellette, G. P., & Sénéchal, M. (2008). A window into early literacy: Exploring the cognitive and linguistic underpinnings of invented spelling. *Scientific Studies of Reading*, *12*(2), 195–219. doi:10.1080/10888430801917324

Ouellette, G., & Sénéchal, M. (2017). Invented spelling in kindergarten as a predictor of reading and spelling in grade 1: A new pathway to literacy, or just the same road, less known? *Developmental Psychology*, *53*(1), 77–88. doi:10.1037/dev0000179

Ouellette, G., Sénéchal, M., & August, J. (2008). Pathways to literacy: A study of invented spelling and its role in learning to read. *Child Development*, *79*(4), 899–913.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315. doi:10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*, 106–151. doi:10.1016/j.cogpsych.2010.04.001

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. doi:10.1037/0033-295X.103.1.56

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www. R-project.org/

Rastle, K. (2019). EPS mid-career prize lecture 2017: Writing systems, reading, and language. *Quarterly Journal of Experimental Psychology*, 1–16. doi:10.1177/1747021819829696

Rodd, J. (2019). How to maintain data quality when you can't see your participants. Retrieved from https://www.psychologicalscience.org/observer/ how-to-maintain-data-quality-when-you-cant-see-your-participants

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. doi:10.1068/p5117

Schad, D. J., Hohenstein, S., Vasishth, S., & Kliegl, R. (2018). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *arXiv Preprint arXiv:1807.10451.*

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174. Retrieved from www.bps.org.uk

Steffensen, M. S., Reynolds, R. E., McClure, E., & Guthrie, L. F. (1982). Black English Vernacular and reading comprehension: A cloze study of third, sixth, and ninth graders. *Journal of Reading Behavior*, *14*(3), 285–298.

Taylor, J. S. H., Davis, M. H., & Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. *Journal of Experimental Psychology: General*, *146*(6), 826–858. doi:10.1037/xge0000301

Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: An artificial orthography paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 60–76. doi:10.1037/a0020126

Terry, N. P., & Scarborough, H. S. (2011). The phonological hypothesis as a valuable framework for studying the relation of dialect variation to early reading skills. In S. A. Brady, D. Braze, & C. A. Fowler (Eds.), *Explaining individual differences in reading: Theory and evidence* (pp. 97–117). New York, NY: Taylor & Francis.

Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading? Cognate effects in a sentence context. *Psychological Science*, *20*(8), 923–927.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*(August), 151–175. doi:10.1016/j.jml.2018.07.004

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels Artificial Character Sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods*, *49*(6), 2093–2112. doi:10.3758/s13428-016-0844-8

Washington, J. A., Branum-Martin, L., Sun, C., & Lee-James, R. (2018). The impact of dialect density on the growth of language and reading in African American children. *Language, Speech, and Hearing Services in Schools*, *49*(2), 232-247.

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Williams, G. P., Panayotov, N. & Kempe, V. (2019, November 20). Literacy Learning in Situations of Dialect Exposure using the Artificial Literacy Learning Paradigm. Retrieved from osf.io/5mtdj.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

Zhu, H. (2019). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Appendix A

Graphemes used to render phonemes in all experiments

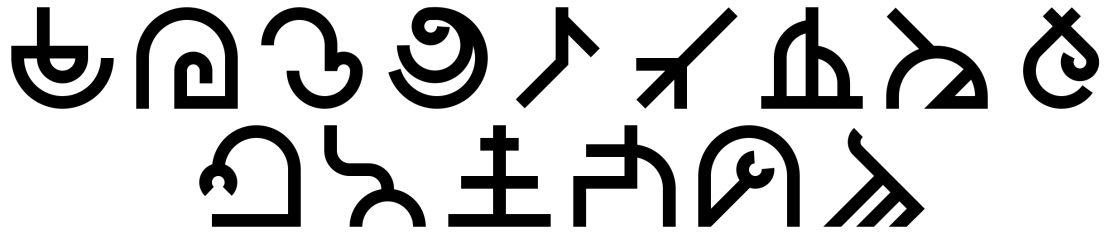

Figure A1: Invented graphemes used to represent each phoneme in all experiments. Note: The final two graphemes were created but not used in these experiments. To prevent participants from memorising the novel graphemes based on resemblance to known graphemes we controlled for similarity to characters of extant writing systems by comparing each invented grapheme against the database of 11,817 characters (excluding Chinese, Korean, and Japanese) on the Shapecatcher website (Milde, 2011). If visual inspection indicated a resemblance, we modified the grapheme to minimise that resemblance.

Appendix B

The Gruffalo books

The Gruffalo.

- The Doric Gruffalo (translated by Sheena Blackhall)
- Thi Dundee Gruffalo (translated by Matthew Fitt)
- The Glasgow Gruffalo (translated by Elaine C. Smith)
- The Gruffalo in Scots (translated by James Robertson)

The Gruffalo's Child.

- The Doric Gruffalo's Bairn (translated by Sheena Blackhall)
- Thi Dundee Gruffalo's Bairn (translated by Matthew Fitt)
- The Gruffalo's Wean (Scots; translated by James Robertson)

Note: The Gruffalo's Child was not available in Glaswegian at the time of this corpus analysis.

Appendix C

List of words and their variants

Word list used in all experiments. Experiment 1, 2b, and 3 used the inconsistent spellings, while Experiment 2a used the consistent spellings.

	Spelling		Pronunciation	
	Consistent	Inconsistent	Non-Contrastive	Contrastive
Training words	nEsk	nEsk	nEsk	nisx
	skEfi	skEfi	skEfi	sxifi
	blEkus	bnEkus	blEkus	blixus
	flEsOd	flEsOd	flEsOd	flisO
	nEf	nEf	nEf	nif
	bEsmi	bEsmi	bEsmi	bismi
	nal	nal	nal	nOl
	daf	daf	daf	dOf
	blaf	bnaf	blaf	blOf
	balf	balf	balf	bOlf
	dasmu	dasmu	dasmu	dOsmu
	smadu	smadu	smadu	smOdu
	kublE	kubnE	kublE	xublE
	slOku	fnOku	slOku	slOxu
	snid	fnid	snid	sni
	fub	fub	fub	
	mif	mif	mif	
	lOm	lOm	lOm	
	snOf	fnOf	snOf	
	blim	bnim	blim	
	flOb	flOb	flOb	
	mOls	mOls	mOls	
	fOns	fOns	fOns	
	nifs	nifs	nifs	
	nOflE	nOflE	nOflE	
	dEsna	dEfna	dEsna	
	smiba	smiba	smiba	
	flidu	flidu	flidu	
	snibOl	fnibOl	snibOl	
	slinab	fninab	slinab	
Testing words	mab	mab	mab	
	skub	skub	skub	
	klEb	klEb	klEb	
	dOlk	dOlk	dOlk	
	suld	suld	suld	

dikla	dikla	dikla
luskO	luskO	luskO
klufE	klufE	klufE
klOda	klOda	klOda
skOnEf	skOnEf	skOnEf
klusim	klusim	klusim
flabun	flabun	flabun

Appendix D

Images used in all experiments:

We selected seven objects from the six categories listed below, resulting in a total of 42 pictures. From each category, we selected images with the highest familiarity scores based on subjective ratings from Rossion and Pourtois (2004), avoiding any items with unclear or incomplete features or those that were deemed to be too similar to another image (e.g. finger and toe) by removing the item with the lower familiarity score and replacing it with the item with the next highest familiarity score in that category (e.g. replacing toe with ear, in this instance).

1. Body part: finger, foot, eye, hand, nose, arm, ear.

2. Furniture and kitchen utensils: chair, glass, bed, fork, spoon, pot, desk.

3. Household objects, tools, and instruments: television, toothbrush, book, pen, refrigerator, watch, pencil.

4. Food and clothing: pants, socks, shirt, sweater, apple, tomato, potato.

5. Buildings, building features, and vehicles: door, house, window, car, doorknob, truck, bicycle.

6. Animals and plants: tree, dog, cat, flower, rabbit, duck, chicken.

The subset of pictures and their associated norms are provided in the supplemental material at https://osf.io/5mtdj/.

Appendix E: Mean proportion of response types for Experiments 1 (panel A), 2a (panel B), 2b (panel C) and 3 (panel D). Response types are: Correct (e.g. target: *kuble* – response: *kuble*); Dialect Word Match: the dialect variant is produced in response to the corresponding standard contrastive word (e.g. target: *kuble* – response: *xuble*); Dialect Word Mismatch: a dialect variant is produced in response to another standard contrastive word (e.g. target: *skefi* – response: *xuble*); Standard Word Mismatch: a standard word is produced in response to another standard word (e.g. target: *skefi* – response: *kuble*), Other Mismatch: any other error that was not part of the response set.